THE ASYMPTOTIC RATE OF THE LENGTH OF THE LONGEST SIGNIFICANT CHAIN WITH GOOD CONTINUATION IN BERNOULLI NET AND ITS APPLICATIONS IN FILAMENTARY DETECTION

A Thesis
Presented to
The Academic Faculty

by

Kai Ni

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Mathematics

Georgia Institute of Technology
May 2013

THE ASYMPTOTIC RATE OF THE LENGTH OF THE LONGEST SIGNIFICANT CHAIN WITH GOOD CONTINUATION IN BERNOULLI NET AND ITS APPLICATIONS IN FILAMENTARY DETECTION

Approved by:

Professor Valdimir Koltchinskii,
Advisor, Committee Chair
School of Mathematics
Georgia Institute of Technology

Professor Xiaoming Huo, Advisor
School of Industrial and Systems
Engineering
Georgia Institute of Technology

Professor Liang Peng
School of Mathematics
Georgia Institute of Technology

Professor Yajun Mei
School of Industrial and Systems
Engineering
Georgia Institute of Technology

Professor Karim Lounici
School of Mathematics
Georgia Institute of Technology

Date Approved: March 26 2013

To my parents, parents in-law and my wife Cui Sun,

for their support and love.

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my advisor Professor Vladimir Koltchinskii for his continuous support , patience and encouragement. Throughout my Ph.D. studies, he has provided knowledgeable mentorship and his sincere suggestions have inspired me, not only in mathematics, but also in becoming a better person.

I would like to specially thank my co-advisor Professor Xiaoming Huo in the School of ISyE for introducing me the interesting projects on image detection. I am sincerely grateful for his constant care and advisement which help me going through the hard times. His discussion with me on our projects inspires me a lot in this applied area.

I would like to thank Professors Vladimir Koltchinskii, Xiaoming Huo, Yajun Mei, Liang Peng and Karim Lounici for being members of my committee and their careful reading of my thesis. I would like to convey my sincere thanks to Professors Christopher Heil, Ionel Popescu, Vladimir Koltchinskii and Ming Yuan for their insightful instruction and inspiring lectures on analysis, probability and statistics. I am also grateful to Professor Haomin Zhou, Jeff Geronimo and Christian Houdre for teaching me wavelets and probability and supporting me as well during the early stage of my research.

It is my pleasure to thank Professor Luca Dieci for recruiting me to Georgia Tech and providing necessary help. I would like to express my gratitude to Cathy Jacobson for her help on my language skills throughout the years. Thank Klara Grodzinsky for her continuous help in my teaching and being my mentor. I am very grateful to Professors Craig Tovey, Santanu Dey, Seong-Hee Kim, Yajun Mei, and Sigrun Andradottir for teaching me courses in optimization, applied probability and

statistics. I would like to extend my gratitude to Professors Anton Leykin, Wing Li, Michael Loss, Heinrich Matzinger and Howie Weiss for giving me enjoyable time in my teaching and grading work.

Thank Ms. Karen Hinds, Ms. Jan Lewis, Ms. Christy Dalton, Ms. Genola Turner, Ms. Sharon McDowell and the IT group for their everyday support.

I am very indebted to many of my colleagues and friends for their helpful discussions and providing a stimulating and fun environment which help me grow during the years. I am especially grateful to my spiritual teachers Daniel Wu and Mark Lin, and the peace fellowship. Many thanks to the friends in the school which help me both in my research and daily life, especially Ke Yin, Jun Lu, Jinyong Ma, Stas Minsker, Ruoting Gong, Yun Gong, Allen Hoffmeyer, Huy Huynh, Linwei Xin, Huijun Feng, Yongfeng Li, Jie Ma, Yao Li and Tianjun Ye.

Lastly and most importantly, I wish to thank my parents, Jianhua Ni and Jiandan Chen, for giving birth and love to me in the first place, my parents in law, Wei Zhou and Qiqin Sun and my wife, Cui Sun, for their love and encouragement. Without their continuous support, it would have been impossible to achieve my goals. To them, I dedicate this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Constantly advanced imaging technology and better software and hardware lead to demands and wishes to use digital images as tools for evaluation and analysis. In most applications, data or images collected by standard sensors such as cameras and radars are analyzed for the detection and recognition of the targets. This thesis is devoted to the detectability of an inhomogeneous region possibly embedded in a noisy environment. It presents models and algorithms using the theory of the longest significant run and percolation; and it analyzes the computational results.

Given a positive integer C, we consider the length of the significant nodes in a chain with good continuation, i.e.,

$$\{(i, j_0), (i+1, j_1), \ldots, (i+k, j_\ell), |j_k - j_{k-1}| \leq C, k = 1, \ldots, \ell\},$$

in a lattice of m rows and n columns of independent nodes and each node is significant independently with probability p. Inspired by the percolation theory, we first analyze the problem in a tree based model

$$\{(i, j) \in \mathbb{Z}^2 : i \geq 0, -iC \leq j \leq iC\}.$$

We give the critical probability and find the decay rate of the probability of having a significant run with length k starting at the origin. Applying the results back to our m and n lattice of nodes, we find the asymptotic rate of the length of the significant run which can be powerfully used in the area of image detection. Examples are detection of filamentary structures in a background of uniform random points in [4] and target tracking problem in [59]. We set the threshold for the rejection region in these problems so that the false positives diminish quickly as we have more samples.

Inspired by the convex set detection in [40], we also give a fast $O(n \log n)$ and near

optimal algorithm to detect a possibly inhomogeneous chain with good continuation in an image of size m-by-n pixels with white noise. We analyze the length of the longest significant chain after thresholding each pixel and consider the statistics over all significant chains. Such a strategy significantly reduces the complexity of the algorithm. The false positives are eliminated as the number of pixels increases. This extends the existing detection method related to the detection of inhomogeneous line segment in [5].

CHAPTER I

INTRODUCTION

1.1 Overview

In application of image detection problems, one class of questions is to determine whether or not some filamentary structures are present in the noisy picture. One approach for this type of detection problems works as follows. At localized batches, hypothesis testing is run to determine whether this batch may overlap with the underlying structure. The hypothesis testing is run while the batch scans through the entire image. The intuition is that if there is an embedded structure, then the significant test results must be clustered around the underlying structure. The difficulty comes from the fact that there will be many false positives among these tests.

We want to take advantage of the fact that the false positive testing results are not clustered, in relative to those that overlap with the underlying feature. Our percolation analysis is motivated by the above phenomenon.

Suppose we have an m-by-n array of nodes. A Bernoulli random variable $X_{i,j}$ is associated with each node (i, j) such that if $X_{i,j} = 1$ then the node is significant (or open); otherwise, insignificant (or closed). However, we suspect that there is a sequence of nodes, with unknown location or orientation, open or closed with a different probability $p_1 > p$. In [11], it is shown that the length of the longest significant run, denoted by $|L_0(m, n)|$ throughout the thesis, has the following asymptotic rate of Erdös-Rényi type (See [8])

$$\lim_{n\to\infty} \frac{|L_0(m,n)|}{\log_{1/\rho(m,p)} n} = 1 \quad \text{almost surely,} \tag{1.1.1}$$

where $\rho(m, p)$ is a constant depending on m and p and also the structure of the model. However the limitation of (1.1.1) is that m is always fixed. Our work extends the

previous work to derive the convergence rate of the length of the longest significant run in the inflating model i.e., $m \to \infty$ and $n \to \infty$ simultaneously. Our theory is related to the percolation theory in which we will introduce the critical probability p_c and divide our theory into $p > p_c$ phase and $p < p_c$ phase. For percolation theory, books by Grimmett [36] and Bollobás [10] are good references. Durrett [22] systematically studies an oriented site percolation model, which is similar to the model in this paper. See also the references therein.

Applications of the aforementioned can be the following:

- Detection of filamentary structures in a background of uniform random points in [4]. We are given N points that might be uniformly distributed in the unit square $[0, 1]^2$. We wish to test whether the set, although mostly consisting of the uniformly scattered points, also contains a small fraction ϵ_N of points sampled from some (unknown a priori) curve with C^α norm bounded by β. See also [25] for a more general case.

- Target tracking problem in [59]. Suppose we have an infrared staring array. A distant moving object will create, upon lengthy exposure, an image of a very faint track against a noisy background. We want to detect whether there is such a moving object in an noisy image.

- Water quality in a network of streams in [30]. Water quality in a network of streams is assessed by performing a chemical analysis at various locations along the streams. As a result, some locations are marked as problematic. We may view the set of all tested locations as nodes and connect pairs of adjacent nodes located on the same stream, thereby creating a tree. We then assign to each node the value 1 or 0, according to whether the location is problematic or not. One can then imagine that one would like to detect a path (or a family of paths) upstream of a certain sensitive location, in order to trace the existence

of a polluter, or look for the existence of an anomalous path upstream from the root of the system.

Recently Langovoy *et al* [46, 47, 18] employs the theory of percolation and random graph to solve the image detection problem. However, our methods in this paper are different from the classic percolation theory, because the nodes here are not necessarily independent *a priori*.

Our work has three advantages.

1. We can drop the independence assumption among nodes which is the fundamental assumption in the percolation theory.

2. Our work is devoted to researching the asymptotic behavior of the longest left-right significant run in the lattice with a varying m.

3. Our model can be easily adapted to the three or higher dimensional cases with some notations change, though for simplicity, the thesis is mostly written based on a 2-dimensional model.

In practice, our work places a fundamental theory on practical problems involving the length of runs. One direct motivation comes from a statistical detection problem. In [4], the authors proposed a method called the multi-scale significance run algorithm (MSRA) for the detection of curvilinear filaments in noisy images. The main idea is to construct a Bernoulli net. Each node has the value of 1 (significant) or 0 (insignificant). Two nodes are defined as connected if they are neighbors (for example their altitude difference is within C), that is, they can simultaneously cover a curve of interest. The length of the nodes in the longest significant run is used as a test statistic. If the length of the run exceeds a certain threshold, then we conclude that there exists an embedded curve; otherwise, there is no embedded curve. To formulate this as a well-defined probability problem, we test the null hypothesis of

a constant success probability p against the alternative hypothesis that some nodes, being on a filament with unknown location and length, have a greater probability of success $p_1 > p$. Under the alternative, the length of the longest significant chain, $|L_0(m,n)|$, is more likely to exceed (i.e., be greater than) a threshold, which, under the null hypothesis, cannot be exceeded. In the approach of [4] the values of these parameters can be chosen for testing. The question is how to choose these parameters so that the power of the test can be maximized. This becomes a design issue. The relation between $|L_0(m,n)|$ and other parameters must be understood. The choice of parameters in the approach of [4] is sufficient to guarantee a proof of asymptotic optimality; Our research systematically searches the relation between $|L_0(m,n)|$ and these parameters.

In [11] the authors show that $\rho(m,p)$ in (1.1.1), which is the limit of conditional probability $\rho_n(m,p)$, lies in $(0,1)$ as $n \to \infty$. Let $\mathcal{A}_{c_1,c_2,\delta_1,\delta_2}$ be the following set

$$\{(m,n) : c_1 n^{1+\delta_1} \leq m \leq c_2 \exp[n(\phi(p) - \delta_2)]\}. \tag{1.1.2}$$

The set $\mathcal{A}_{c_1,c_2,\delta_1,\delta_2}$ essentially states that as the column number n increases, m increases faster than any linear growth of n and slower than some exponential growth of n. In our work, in the case of $p < p_c$, as $m \to \infty$, $n \to \infty$ and $(m,n) \in \mathcal{A}_{c_1,c_2,\delta_1,\delta_2}$, we have

$$\rho_n(m,p) \to exp\{-\phi(p)\};$$

and

$$|L_0(m,n)| = \log(mn)/\phi(p) + o_p(1),$$

where $\phi(p)$ is a positive function and will be defined in (2.2.6).

Applying our theory to the multi-scale detection method in [4], we describe a multi-scale significant run algorithm that can reliably detect the concentration of data near a smooth curve, without knowing the smoothness information α or β in advance, provided that the portion of points on the curve ϵ_N exceeds $T(\alpha,\beta)N^{1/(1+\alpha)}$.

Our $T(\alpha, \beta)$ is smaller than that in [4] which indicates stronger detection ability using our theory. In the target tracking problem, our method provides a reliable threshold such that the false alarm probability vanishes very quickly as we get more and more sample points.

As another application of the theory, we will also present an image detection model on white noise. Figure 1 (a) contains a chain with good continuation in an image with 10-by-60 pixels; (b)-(d) present the same chain in noisy Gaussian random field in images with different elevated means. The detectability problem is to ask: when is the chain detectable and what is the order of complexity for the detecting algorithms? We consider the following statistical formulation: the intensity at each pixel follows a normal distribution. Inside the chain, the normal mean μ is a positive constant such as in Figure 1 (b) $\mu = 1.0$, (c) $\mu = 2.5$ and (d) $\mu = 4.0$; while outside the chain, the normal mean is 0. In (b) and (c), the chain can hardly be observed by eyes while in (d), the chain is clearly visible.

According to our result, a complete enumeration of all possible chains with good continuation in an image with m-by-n pixels is not a good choice because the cardinality of such an enumeration will be approximately $O(e^n)$. However, it is doable if we give a constraint on the length of chains, which reduces the number of chains under consideration to $O(n^2)$. For chains with length larger than this constraint, we implement a detecting method based on the longest significant chain in an m-by-n array as in [11]. We first use a threshold to classify each pixel as either significant or insignificant based on pixel's intensity. In [11], the authors present an asymptotic rate of the length of the longest significant chain in a Bernoulli net. When applying to our problem with chains consisting of only significant nodes, this technique again significantly reduces the number of chains under consideration to a polynomial formula of n. We implement our method and find in most cases, the detectable mean lies in between 1 and 2. This result is much better than the detectability by human

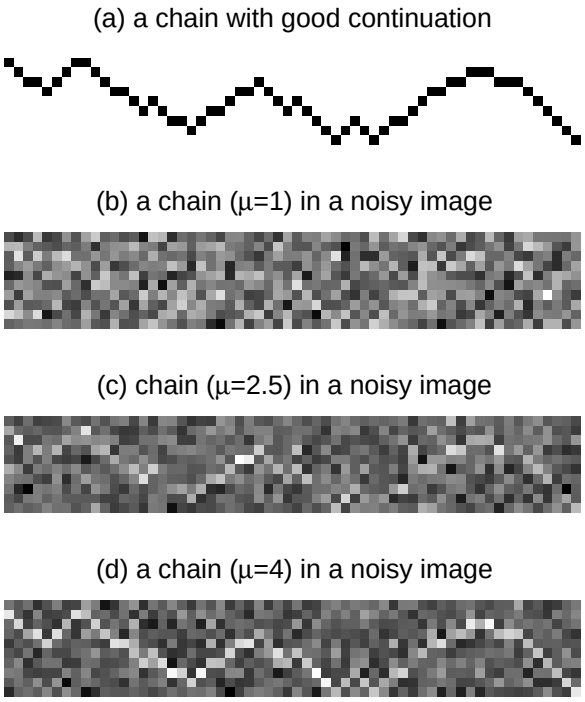

Figure 1: A chain with good continuation (a) and chains embedded in noisy images with mean 1 in (b); 2.5 in (c); and 4.0 in (d).

.

eyes, in which people can hardly tell the embedded chain with confidence when $\mu < 3$ as shown in Figure 1.

Our algorithm has three advantages. First, the algorithm has very low order of complexity $O(n \log n)$ to detect the embedded chain. Note that there are $O(e^n)$ possible chains under consideration, so our algorithm is very fast. Second, our detection algorithm is asymptotically powerful which means as the size of the noisy image becomes larger and larger, the detection errors (type-I error and type-II error) go to zero. Third, even if the length of the embedded inhomogeneous chain is short as $O(\log n)$, the minimum detectable elevated mean in the embedded chain is almost always a half of what can be detected by eyes. Thus, our algorithm is good in terms of stability.

The rest of the thesis is organized as follows. In Section 1.2, we present a review of the related and existing work. In Chapter 2, we present a pseudo-tree model and study the critical probability and its reliability problems. In Chapter 3, we summarize the previous work of the Bernoulli net and give the extensions beyond the fixed number of rows. Chapter 4 presents possible applications of the longest run method in image detection problems. In Chapter 5, a fast and near optimal algorithm is given to detect inhomogeneous chains of low strength in an image of white noise; we theoretically give the lowest possible detectable strength of the inhomogeneous region using this algorithm and show the stimulation and numeric studies.

1.2 Literature Review

There is a plethora of available statistical methods that can, in principle, be used for filaments detection and estimation. These include: Principle curves in [37], [41], [65] and [66]; nonparametric, penalized, maximum likelihood in [71]; parametric models in [67]; manifold learning techniques in [61], [70] and [39]; gradient based methods in [54] and [35]; methods from computational geometry in [19], [49] and [12]; faint line segment detection in [14]; Ship Wakes "V" shape detection against a highly cluttered background in [13] and underlying curvilinear structure in [4], [59] and [5]. See also [46], [47] and [18] for the applications of the percolation theory in this area.

There is a multitude of applications for which our model is relevant. Examples include the detection of hazardous materials [38] and target tracking [16] in sensor networks [15], and disease outbreak detection [58]. Pixels in digital images are also sensors so that many other examples can be found in the literature on image processing such as road tracking [34] and fire prevention using satellite imagery [17] and the detection of tumors in medical imaging [52].

The generalized likelihood ratio test, which is known as the scan statistic in spatial statistics [42, 43], is by far the most popular method in practice and is given different

names in different fields. Most of the methods related to scan statistic assume that the clusters are in some parametric family such as circular [44], elliptical [1, 50] or, more generally, deformable templates [2] while others do not assume explicit shapes [20, 51, 69], which leads to nonparametric models.

We consider a nonparametric method based on the percolative properties of the network. The most basic approach is based on the size of the largest significant chain of the graph after removing the nodes whose values fall under a given threshold. If the graph is a one-dimensional lattice, after thresholding this corresponds to the test based on the longest run [9], which [11] adapts for path detection in a thin band. This test is studied in a series of papers such as [46, 45] under the name of maximum cluster test. More sophisticated is the upper level set scan statistic of [32, 31, 33]. In its basic form, it scans over the connected components of the graph after thresholding.

The task of detection in networks is critical for an increasing number of applications, for example, in surveillance and environment monitoring. Some of these applications are:

- Detection in sensor networks. Sensor networks offer a more flexible, decentralized alternative and are considered for the detection of radioactive, biological or chemical materials [64, 72]. Sensor networks are also used in other target tracking settings [16].

- Disease outbreak detection. Some specific information networks are used, with surveillance systems now incorporating data from hospital emergency visits, ambulance dispatch calls and pharmacy sales of over-the-counter drugs [58, 60].

- Virus detection in a computer network. Diseases affect computers in the form of viruses and worms spreading from host to host in a computer network [68].

- Detection from field measurements. The objective of assessing the water quality is to determine whether there are regions of low biological integrity based on

the collected data, and to identify these regions [31].

CHAPTER II

PSEUDO-TREE MODEL

2.1 Model Introduction

In this section, we first present a model which has some similarity to a *regular* or *complete-tree* model ([3, 24]). Consider, for example, the lattice with nodes of the form

$$V = \{(i, j) \in \mathbb{Z}^2 : -iC \leq j \leq iC, i \geq 0\} \tag{2.1.3}$$

and oriented edges $(i, j) \to (i+1, j+s)$, where $|s| \leq C$. We call $(0, 0)$ the origin of the graph and sometimes use 0 to denote the origin. Let $Y_{i,j}$ be the i.i.d. Bernoulli(p) state variables corresponding to node (i, j). If $Y_{i,j} = 1$, we say the node (i, j) is significant and we are interested in the length of significant runs starting at the origin. See Figure 2 for a sketch of the model.

Note that even though the number of runs of length k in Pseudo-tree model and the regular tree model with $2C+1$ descendants are the same (both equal $(2C+1)^{k-1}$), the numbers of nodes are considerably different in the first k columns—about k^2C for the former and about $(2C + 1)^k$ for the regular tree .

Let p_c denote the critical probability for the site percolation in the Pseudo-tree model, denoted as the supremum over all $p \in (0, 1)$ such that the size of the significant run at the origin is finite with probability one. By our knowledge, this model has not been fully studied yet and we will elaborate some results in the next section. Analogous to the model presented here, recent papers ([24, 6]) have studied the oriented and non-oriented significant clusters or runs in a regular lattice.

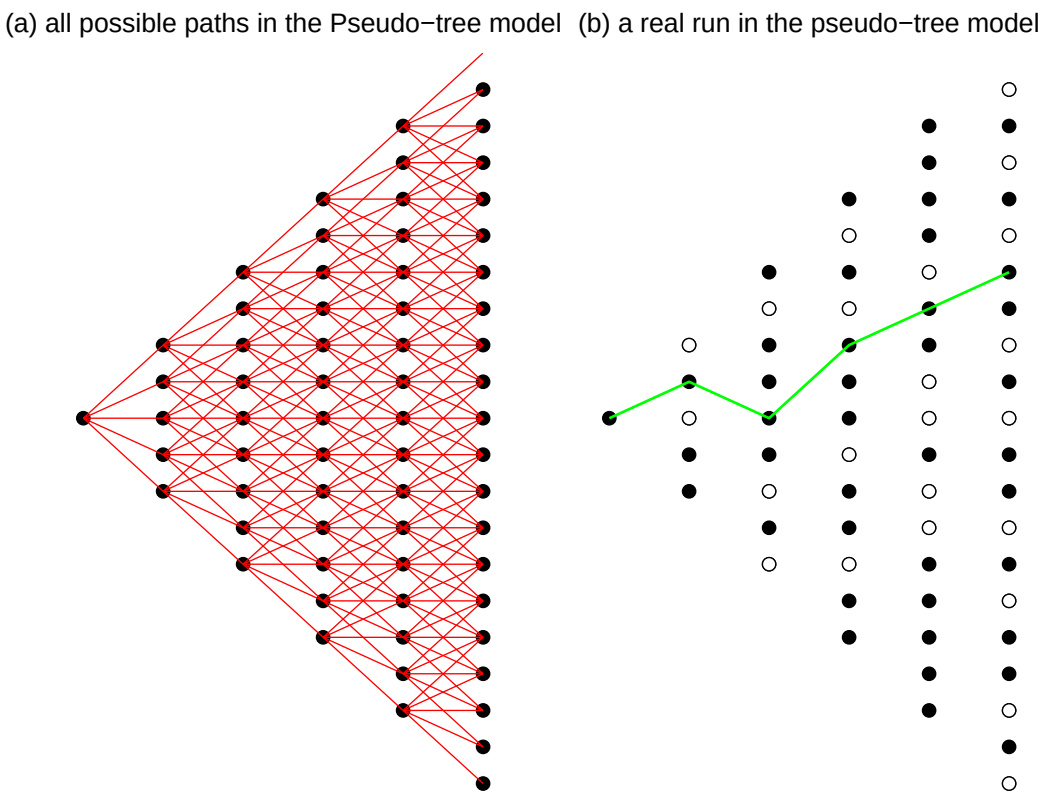

Figure 2: A sketch of pseudo-tree model with the connectivity constraint $C = 2$. (a) gives all the possible edges in the model. In (b) solid nodes are significant. The green path shows a possible real run in the pseudo-tree model

.

2.2 Results

In this section, we give some results about the significant runs in Pseudo-tree model V presented in (2.1.3). The difference between the *Pseudo-tree* and *Regular-tree* model is that the number of nodes in the former grows linearly with the depth, as opposed to grow exponentially with the depth in the latter. Besides, in *Pseudo-tree* model, different runs may share the same edges and therefore the behaviors of distinct runs here are quite correlated.

2.2.1 Notation

We shall introduce some notation. Observe that there is only one node in the 0-th column, namely the origin $(0,0)$ and there are $2kC+1$ nodes in the k-th column, namely the nodes $(k,-kC),\ldots,(k,0),\ldots,(k,kC)$. For $k \in \mathbb{Z}^+$, let $B(k) = \{(k,-kC),\ldots,(k,0),\ldots,(k,kC)\}$ be the set of nodes in k-th column in V.

Let $\theta_k(p)$ denote the probability that $(0,0)$ is connectible to the $(k-1)$th column by a significant run, i.e., $\theta_k(p) = \mathbb{P}_p((0,0) \leftrightarrow B(k-1))$. In other words, $\theta_k(p)$ is the probability that there is a significant run of length at least k starting at the origin. Given any $x = (x_1,x_2) \in \mathbb{Z}^2$, let $\theta_k^x(p)$ be the probability that x connects the (x_1+k-1)-th column with a significant chain. It is easy to see that $\theta_k^x(p)$ does not depend on the status of the nodes before the x_1-th column and $\theta_k^x(p) = \mathbb{P}_p(\{x \leftrightarrow B(x_1+k-1)\}) = \theta_k(p)$. Because $\theta_k(p)$ only involves finitely many nodes, one can easily see that $\theta_k(p)$ is a continuous function of $p \in [0,1]$. Throughout the dissertation, we will use sometimes n as a subscript instead of k.

2.2.2 Critical Probability

Given the above, we have some properties

- $\theta_{k_1}(p) \le \theta_{k_2}(p)$ if $k_1 \ge k_2$ which implies $\theta(p) \equiv \lim_{k\to\infty} \theta_k(p)$ exists;
- $\theta_k(0) = 0$ and $\theta_k(1) = 1$ for any $k \ge 1$ which implies $\theta(0) = 0, \theta(1) = 1$;
- $\theta_k(p)$ and $\theta(p)$ are nondecreasing with respect to p.

Thus $\theta(p)$ is the probability that there is a significant run in V starting from the origin and heading towards right forever when the probability of a node to be open is p. In light of this, we define p_c to be the critical probability, i.e.,

$$p_c \equiv \sup\{p \in [0,1] : \theta(p) = 0\}.$$

So p_c is the critical probability, above which it is possible to have an infinite significant run starting from any node in *Pseudo-tree* model.

Recall that in the r-regular tree model, the critical probability $p_c = 1/r$. Our first result shows that in the *Pseudo-tree* model, the critical probability is no smaller than $1/r$, where $r = 2C + 1$ (See [10]).

Theorem 2.2.1. *The critical probability p_c of the Pseudo-tree model is $\geq \frac{1}{2C+1}$.*

In the beam-let model of [4], each node is connectible to 81 nodes in the next column. Thus this theorem explains the reason that the authors there took the membership threshold N^* such that $p = \mathbb{P}(\text{Poisson}(2) > N^*) = \frac{p_0}{81}$ for some $p_0 \in (0, 1)$. The proof of Theorem 2.2.1 requires the following definition and lemma. See [36].

Definition 2.2.2. *Let V be the set of nodes in the pseudo-tree and we take the sample space as*

$$\Omega = \prod_{v \in V} \{0, 1\}.$$

We take $\mathcal{F}$ to be the σ-field os subsets of Ω generated by the finite-dimensional cylinders. We say an event $A \in \mathcal{F}$ is increasing if the indicator function of A satisfies $I_A(X_1) \leq I_A(X_2)$ whenever $X_1 \leq X_2$, where X_1, X_2 are two realizations on V, i.e., $X_1 : V \to \{0, 1\}$, where $X_1(i, j) = 1$ if the node (i, j) is significant and $X_1(i, j) = 0$ otherwise and X_2 has the same definition. Analogously, we say A decreasing set if $\overline{A}$, the complement of A, is increasing.

Lemma 2.2.3. *(FKG Inequality) If A and B are both increasing (or both decreasing) events in the lattice, then we have $\mathbb{P}(A \bigcap B) \geq \mathbb{P}(A)\mathbb{P}(B)$.*

The significant edge version of this lemma can be found in Section 2.2 of [36]. The intuition behind this lemma is that if there is an open path joining vertex u to vertex v, then it becomes more likely that there is an open path joining vertex x to vertex y than without a path from u to v. Replacing edge by node in the proof in [36], the significant node version can be shown analogously.

Proof of Theorem 2.2.1. Recall that $\theta_k(p) = \mathbb{P}_p(0 \leftrightarrow B(k-1))$. The event $\{0 \leftrightarrow B(k)\}$ happens if and only if there is an open node $x \in B(1)$, such that the origin $(0,0)$ is open and the event $\{x \leftrightarrow B(k)\}$ occurs. Of course, $\text{card}(B(1)) = 2C+1$. Therefore we have,

$$\{0 \leftrightarrow B(k)\} = \{ \bigcup_{\{x \in B(1)\}} ((0 \leftrightarrow x) \cap (x \leftrightarrow B(k)))\}$$

This implies that

$$\begin{aligned} 1 - \mathbb{P}_p(0 \leftrightarrow B(k)) &= \mathbb{P}_p(\bigcap_{\{x \in B(1)\}} \overline{(0 \leftrightarrow x) \bigcap (x \leftrightarrow B(k)))} \\ &\geq \prod_{\{x \in B(1)\}} \mathbb{P}_p(\overline{(0 \leftrightarrow x) \bigcap (x \leftrightarrow B(k)))} \\ &= (1 - p\theta_k(p))^{(2C+1)}, \end{aligned} \tag{2.2.4}$$

where the inequality is due to Lemma 2.2.3 and the fact that $\{0 \leftrightarrow x\} \cap \{x \leftrightarrow B(k)\}$ is an increasing event and that $\mathbb{P}_p(x \leftrightarrow B(k)) = \mathbb{P}_p(0 \leftrightarrow B(k-1))$ for any given $x \in B(1)$.

So by (2.2.4), we have that

$$\begin{aligned} \theta_{k+1}(p) &= \mathbb{P}_p(0 \leftrightarrow B(k)) \\ &\leq 1 - (1 - p\mathbb{P}_p(0 \leftrightarrow B_{k-1}))^{(2C+1)} \\ &= 1 - (1 - p\theta_k(p))^{(2C+1)}. \end{aligned}$$

Given this, we investigate the function

$$f(x) = 1 - (1 - px)^k$$

where $k \in \mathbb{Z}^+$. We have

$$f'(x) = kp(1-px)^{k-1} > 0 \quad \& \quad f''(x) = -k(k-1)p^2(1-px)^{k-2} < 0, \forall x \in (0,1).$$

So the function $f(x)$ is always strictly increasing and concave down and $f(0) = 0$. Besides, one can see that $f'(0) = kp$ and from this we have $f(x)$ is always under the

line $y = x$ if $p < \frac{1}{k}$. Let x_0 be an arbitrary number in $(0, 1)$ and generate a sequence $\{x_n\}_{n\geq 0}$ such that $x_{n+1} = f(x_n)$ for $n \geq 0$. Since $f(x) < x$ when $x \in (0, 1)$ and $p < \frac{1}{k}$, the sequence $\{x_n\}_{n\geq 0}$ is strictly decreasing. On the other hand, it is easy to see that $x_n \geq 0$ for any $n \geq 0$. Because a bounded decreasing sequence must have a limit, we have that

$$0 \leq x^* \equiv \lim_{n\to\infty} x_n.$$

By the continuity of $f(x)$, one can easily see that

$$f(x^*) = \lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} x_{n+1} = x^*.$$

Since $f(x) < x$ on $(0, 1)$, it is obvious that the limit of the sequence $\{x_n\}_{n\geq 0}$ is 0, i.e., $x^* = 0$ for any starting point $x_0 \in (0, 1)$.

Hence when $p < \frac{1}{(2C+1)}$, it leads to

$$\theta(p) = \lim_{k\to\infty} \theta_{k+1}(p) = \lim_{k\to\infty} \mathbb{P}(0 \leftrightarrow B(k)) \leq \lim_{k\to\infty} 1 - (1 - p\theta_k(p))^{(2C+1)} = 0.$$

It follows that $p_c \geq \frac{1}{(2C+1)}$. □

2.2.3 Asymptotic rate of $\theta_k(p)$

In this part, we show that under the sub-critical phase $p < p_c$

$$\theta_k(p) = \mathbb{P}_p(0 \leftrightarrow B(k-1)) = O(k \exp\{-k\phi(p)\}),$$

where $\phi(p) > 0$ is a decreasing function of p.

Theorem 2.2.4. *Suppose $0 < p \leq 1$. There exist positive constants σ_1 and σ_2, independent of p, and a unique function $\phi(p)$, such that*

$$\sigma_1 k^{-1} \exp\{-k\phi(p)\} \leq \theta_k(p) \leq \sigma_2 k \exp\{-k\phi(p)\} \tag{2.2.5}$$

for any $k \geq 1$. In particular,

$$\frac{\log \theta_k(p)}{k} \to -\phi(p). \tag{2.2.6}$$

Before proving the theorem, let us state the sub-additivity lemma which can be found in [36].

Lemma 2.2.5. ***Sub-additive limit theorem.*** *If $(x_r : r \geq 1)$ is sub-additive, i.e., $x_{m+n} \leq x_m + x_n$ for all m, n, then $\lambda = \lim_{r\to\infty}\{\frac{x_r}{r}\}$ exists and satisfies $-\infty \leq \lambda < \infty$. Furthermore*

$$\lambda = \inf\{\frac{x_m}{m} : m \geq 1\}$$

and thus $x_m \geq m\lambda$ for all m.

Proof of Theorem 2.2.4. Given a positive integer l, it is not hard to show that

$$\mathrm{card}(B(l)) = (2lC + 1).$$

Since the event $\{0 \leftrightarrow B(l+k)\}$ occurs if and only if there is some $z \in B(l)$ such that both $\{0 \leftrightarrow z\}$ and $\{z \leftrightarrow B(l+k)\}$ occur, we have

$$\begin{aligned}
\theta_{k+l}(p) = \mathbb{P}_p(0 \leftrightarrow B(k+l-1)) &= \mathbb{P}_p(\bigcup_{\{z\in B(l)\}} (\{0 \leftrightarrow z\} \cap \{z \leftrightarrow B(k+l-1)\})) \\
&\leq \sum_{\{z\in B(l)\}} \mathbb{P}_p(\{0 \leftrightarrow z\} \cap \{z \leftrightarrow B(k+l-1)\}) \\
&= \frac{1}{p}\sum_{\{z\in B(l)\}} \mathbb{P}_p(\{0 \leftrightarrow z\})\mathbb{P}_p(\{z \leftrightarrow B(k+l-1)\}) \\
&= \frac{1}{p}\sum_{\{z\in B(l)\}} \mathbb{P}_p(\{0 \leftrightarrow z\})\mathbb{P}_p(\{0 \leftrightarrow B(k-1)\}),
\end{aligned}$$

where the last equality is due to the fact that

$$\mathbb{P}_p(\{z \leftrightarrow B(k+l-1)\}) = \mathbb{P}_p(\{0 \leftrightarrow B(k-1)\}), \forall z \in B(l).$$

Notice that $\mathbb{P}_p(0 \leftrightarrow z) \leq p\theta_l(p)$ for any $z \in B(l)$. We have

$$\theta_{k+l}(p) \leq (2lC + 1)\theta_l(p)\theta_k(p).$$

On the other hand, for any $z \in B(l)$, we have

$$\begin{aligned}
\theta_{k+l}(p) &= \mathbb{P}_p(\{0 \leftrightarrow B(k+l-1)\}) \\
&\geq \mathbb{P}_p(\{0 \leftrightarrow z\} \cap \{z \leftrightarrow B(k+l-1)\}) \\
&= \frac{1}{p}\mathbb{P}_p(\{0 \leftrightarrow z\})\mathbb{P}_p(\{z \leftrightarrow B(k+l-1)\}) \\
&= \frac{1}{p}\mathbb{P}_p(\{0 \leftrightarrow z\})\mathbb{P}_p(\{0 \leftrightarrow B(k-1)\}),
\end{aligned}$$

Notice that

$$\theta_l(p) \leq \frac{1}{p}\mathbb{P}_p(\bigcup_{\{z \in B(l)\}} \{0 \leftrightarrow z\}) \leq \frac{1}{p}\sum_{\{z \in B(l)\}} \mathbb{P}_p(\{0 \leftrightarrow z\}).$$

It follows to have a node $z \in B(l)$ such that

$$\frac{1}{p}\mathbb{P}_p(\{0 \leftrightarrow z\}) \geq \frac{\theta_l(p)}{(2lC+1)}.$$

Thus we have

$$\theta_{k+l}(p) \geq \frac{1}{(2lC+1)}\theta_l(p)\theta_k(p).$$

If we let $g(l) = \log(2lC+1)$, then the inequalities we get so far are:

$$\begin{aligned}
\log(\theta_{k+l}(p)) &\leq \log(\theta_k(p)) + \log(\theta_k(p)) + g(l); \\
\log(\theta_{k+l}(p)) &\geq \log(\theta_k(p)) + \log(\theta_l(p)) - g(l).
\end{aligned}$$

Notice that $g(k+l) - g(k) = \log(1 + \frac{2lC}{2kC+1}) \leq \log 2$ if $l \leq k$. Therefore, we have

$$\begin{aligned}
&\log(\theta_{k+l}(p)) + g(k+l) + \log 2 \\
&\leq \log(\theta_k(p)) + g(k) + \log 2 + \log(\theta_l(p)) + g(l) + \log 2; \qquad (2.2.7)
\end{aligned}$$

$$\begin{aligned}
&\log(\theta_{k+l}(p)) - g(k+l) - \log 2 \\
&\geq \log(\theta_k(p)) - g(k) - \log 2 + \log(\theta_l(p)) - g(l) - \log 2. \qquad (2.2.8)
\end{aligned}$$

Then by Lemma 2.2.5, we have

$$\begin{aligned}\phi(p) &:= \lim_{k\to\infty} -\frac{1}{k}\{\log(\theta_k(p))\} \\ &= \lim_{k\to\infty} -\frac{1}{k}\{\log(\theta_k(p)) + g(k) + \log 2\} \\ &= \lim_{k\to\infty} -\frac{1}{k}\{\log(\theta_k(p)) - g(k) - \log 2\}.\end{aligned}$$

This leads to

$$\log(\theta_k(p)) + g(k) + \log 2 \geq -k\phi(p); \tag{2.2.9}$$

$$-\log(\theta_k(p)) + g(k) + \log 2 \geq k\phi(p) \tag{2.2.10}$$

for all $k \geq 1$. The theorem now follows (2.2.9) and (2.2.10) easily. □

The next corollary gives the limit of $\frac{\theta_k(p)}{\theta_{k-1}(p)}$.

Corollary 2.2.6.

$$\lim_{k\to\infty} \frac{\theta_k(p)}{\theta_{k-1}(p)} = \exp\{-\phi(p)\}. \tag{2.2.11}$$

Proof of Corollary 2.2.6. By inequality (2.2.7), we know that the sequence

$$\{\log(\theta_k(p)) + g(k) + \log 2\}_{k\in\mathbb{N}}$$

is a sub-additive sequence. Thus by Lemma 2.2.5 we have

$$\begin{aligned}-\phi(p) &= \lim_{k\to\infty} \frac{\log(\theta_k(p)) + g(k) + \log 2}{k} \\ &= \inf_{k\in\mathbb{N}} \frac{\log(\theta_k(p)) + g(k) + \log 2}{k}\end{aligned}$$

Therefore, for any $\epsilon_0 > 0$, there exists some large k_0 such that when $k > k_0$, we have

$$-\phi(p) \leq \frac{\log(\theta_k(p))}{k} + \epsilon_0,$$

which leads to

$$\exp(-\phi(p)) \leq (\theta_k(p))^{\frac{1}{k}} \exp(\epsilon_0), \forall k > k_0.$$

By inequality (2.2.8), we know that $\{g(k) + \log 2 - \log(\theta_k(p))\}_{k\in\mathbb{N}}$ is a sub-additive sequence, therefore we have

$$g(k) + \log 2 - \log(\theta_k(p)) \leq g(k-1) + \log 2 - \log(\theta_{k-1}(p)) + g(1) + \log 2 - \log(\theta_1(p)).$$

Divide by k on the left and by $k-1$ on the right. It is easy to see that for any $\epsilon_1 > 0$, there exists some large k_1 such that when $k > k_1$, we have

$$\frac{\log(\theta_k(p))}{k} \geq \frac{\log(\theta_{k-1}(p))}{k-1} - \epsilon_1.$$

It follows that when $k > \max\{k_0, k_1\}$, we have

$$\exp(-\phi(p)) \leq (\theta_k(p))^{\frac{1}{k}} \exp(\epsilon_0) \leq \frac{\theta_k(p)}{\theta_{k-1}(p)} \exp(\epsilon_0 + \epsilon_1/k).$$

By the same technique using (2.2.8) and (2.2.7), we can show that for any ϵ_2, when $k > k_2$ for some large k_2, we have

$$\exp(-\phi(p) + \epsilon_2) \geq (\theta_k(p))^{\frac{1}{k}} \geq \frac{\theta_k(p)}{\theta_{k-1}(p)}.$$

Since ϵ_0, ϵ_1 and ϵ_2 are arbitray, we have that

$$\lim_{k\to\infty} \frac{\theta_k(p)}{\theta_{k-1}(p)} = \exp\{-\phi(p)\}.$$

□

Given Theorem 2.2.4, one may speculate that $\phi(p) \to \infty$ as $p \to 0$ since $\theta(p) = 0$ as $p = 0$ and the theorem merits when $\phi(p) > 0$. We will show $\phi(p)$ has the desired properties as $p < p_c$ in the following corollary.

Corollary 2.2.7. *The function* $\phi(p) := \lim_{k\to\infty} -\frac{\log \theta_k(p)}{k}$ *satisfies the following:*

1. $\phi(p)$ *is a continuous function on* $(0, 1]$*;*

2. $\phi(p)$ *is strictly decreasing on* $(0, p_c)$ *and constantly* 0 *when* $p_c \leq p \leq 1$*;*

3. $\lim_{p\to 0} \phi(p) = \infty$.

Remark 2.2.8. *By observing Corollary 2.2.7, the Theorem 2.2.4 is of no value when* $p \geq p_c$ *because* $\phi(p)$ *is constantly* 0 *in the supercritical phase.*

Figure 3 gives the tendency of $-\frac{\log \theta_k(p)}{k}$ against k for different values of p when $C = 1$.

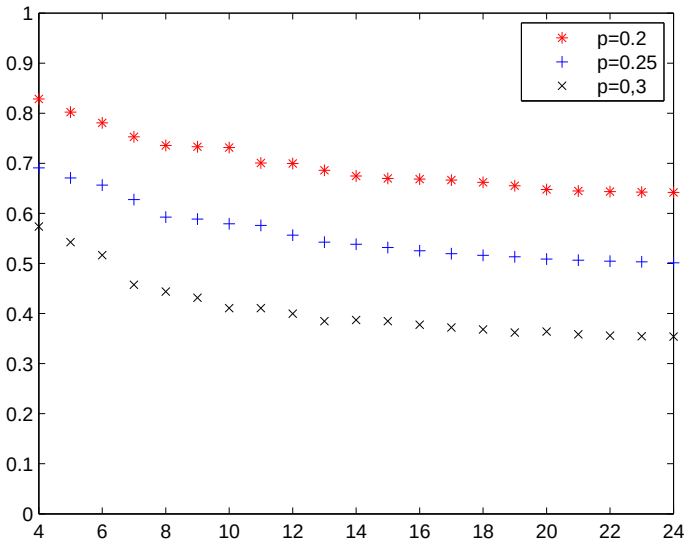

Figure 3: A sketch of simulated result of $-\log \frac{\theta_k(p)}{k}$ against k with p being 0.2, 0.25, 0.3 when $C = 1$

To prove the corollary, we have to introduce the following lemma.

Lemma 2.2.9. *Let* A *be an increasing event which depends on only finitely many nodes of a lattice. Then* $\frac{\log \mathbb{P}_p(A)}{\log p}$ *is a non-increasing function of* p.

The proof of this lemma in bond percolation problem is proved in [36]. For self-completeness, we include the proof of this lemma after Corollary 2.2.7.

Proof of Corollary 2.2.7. It is easy to see $\phi(p) = 0$ if $p > p_c$. Indeed, since

$$\mathbb{P}_p(0 \leftrightarrow B(k)) \geq \theta(p) > 0,$$

it leads to

$$0 \leq \phi(p) = \lim_{k \to \infty} -\frac{\log \mathbb{P}_p(0 \leftrightarrow B(k-1))}{k} \leq \lim_{K \to \infty} -\frac{\log \theta(p)}{k} = 0,$$

when $p > p_c = \frac{1}{(2C+1)}$.

Since $\theta_k(p)$ only depends on the status of finitely many sites, $-\frac{1}{k}\log(\theta_k(p))$ is a continuous function of p for any $k \geq 1$. So it is sufficient to show that $-\frac{1}{k}\log(\theta_k(p))$ converges to $\phi(p)$ uniformly on $(0, 1]$. By (2.2.9) and (2.2.10), we have for any $p \in (0, 1]$

$$\left|\phi(p) + \frac{1}{k}\log(\theta_k(p))\right| \leq \frac{1}{k}(g(k) + \log 2) \to 0 \quad \text{as} \quad k \to \infty$$

which does not depend on p at all. So $\phi(p)$ is a continuous on $(0, 1]$. And it follows the fact that $\phi(p_c) = 0$, since

$$\phi(p_c) = \lim_{p \downarrow p_c} \phi(p) = 0$$

by continuity of $\phi(p)$. To prove the strict monotonicity of $\phi(p)$ when $0 < p < p_c$, we notice that $\{0 \leftrightarrow B(k-1)\}$ is an increasing event which only depends on finitely many edges. Thus we apply the Lemma 2.2.9 to have that

$$\frac{\log \mathbb{P}_a(0 \leftrightarrow B(k-1))}{\log a} \geq \frac{\log \mathbb{P}_b(0 \leftrightarrow B(k-1))}{\log b} \quad \text{if} \quad a \leq b.$$

If we divide the above by k and take the limit as $k \to \infty$, then we have

$$\phi(a) \geq \phi(b)\frac{\log(\frac{1}{a})}{\log(\frac{1}{b})} \quad \text{if} \quad 0 < a \leq b \leq 1.$$

So if $0 < a < b < p_c$, $\phi(a) > \phi(b)$.

To prove that $\lim_{p \to 0} \phi(p) = \infty$, we use $\chi(k)$ and $\chi^*(k)$ to denote the number of all runs and significant runs respectively in the Pseudo-tree model that connect 0 and $B(k-1)$ respectively. It is not hard to see that $\chi(k) = (2C+1)^{k-1}$ and $\mathbb{E}_p(\chi^*(k)) = p^k \times \chi(k)$. Therefore, we have the following

$$\begin{aligned}
\theta_k(p) &= \mathbb{P}_p(0 \leftrightarrow B(k-1)) \\
&= \mathbb{P}_p(\chi^*(k-1) \geq 1) \\
&\leq \mathbb{E}_p(\chi^*(k-1)) \\
&= p^{k-1} \times (2C+1)^{k-1}
\end{aligned}$$

So this will lead to the following fact

$$\begin{aligned}\lim_{k\to\infty} -\frac{\log\theta_k(p)}{k} &\geq \lim_{k\to\infty} -\log(p\times(2C+1))\times\frac{k-1}{k}\\ &= -\log(p\times(2C+1)).\end{aligned}$$

So as $p\to 0$, obviously $\phi(p)\to\infty$. □

Proof of Lemma 2.2.9. For any $0<p_1\leq p_2<1$, we have $\gamma\geq 1$ such that $p_1=p_2^{\gamma}$. Thus it suffices to show that

$$f(p^{\gamma})\leq (f(p))^{\gamma}, \forall p\in(0,1) \quad \text{and} \quad \forall\gamma\geq 1, \tag{2.2.12}$$

where $f(p)=\mathbb{P}_p(A)$, where A is some increasing event which depends on finitely many nodes of the lattice and p is the marginal probability of a node to be significant. Let us show (2.2.12) by induction on the number of nodes on which A depends. Let A_s be the set of nodes that A depends on. If A_s is a singleton, i.e., there is only one node $a\in A_s$, then

$$f(p^{\gamma})=\mathbb{P}_{p^{\gamma}}(A)=(f(p))^{\gamma}=(\mathbb{P}_p(A))^{\gamma}.$$

Suppose now that $k\geq 1$ is such that (2.2.12) is valid whenever $\text{card}\{A_s\}\leq k$, and consider the case that $\text{card}\{A_s\}=k+1$. Let $0<p<1$, $\gamma\geq 1$ and let $z\in A_s$ be one node on which A depends. Let $\eta(z)$ be the indicator of z, i.e, $\eta(z)=1$ if the site z is open and 0 otherwise. Then we have the following

$$\begin{aligned}f(p^{\gamma}) &= \mathbb{P}_{p^{\gamma}}(A\big|\eta(z)=1)p^{\gamma}+\mathbb{P}_{p^{\gamma}}(A\big|\eta(z)=0)(1-p^{\gamma})\\ &\leq \mathbb{P}_p(A\big|\eta(z)=1)^{\gamma}p^{\gamma}+\mathbb{P}_p(A\big|\eta(z)=0)^{\gamma}(1-p^{\gamma})\end{aligned}$$

by the induction hypothesis. The following inequality is not hard to prove and we will prove it later

$$x^{\gamma}p^{\gamma}+y^{\gamma}(1-p^{\gamma})\leq\{xp+y(1-p)\}^{\gamma}, \tag{2.2.13}$$

when $x\geq y\geq 0, 0<p<1, \gamma\geq 1$. It is not difficult to see that

$$x=\mathbb{P}_p(A\big|\eta(z)=1)\geq\mathbb{P}_p(A\big|\eta(z)=0)=y$$

since A is an increasing event. Given (2.2.13), we note that

$$f(p^\gamma) \leq \{\mathbb{P}_p(A\big|\eta(z)=1)p + \mathbb{P}_p(A\big|\eta(z)=0)(1-p)\}^\gamma = f(p)^\gamma.$$

To see that (2.2.13), check that equality holds when $x = y \geq 0$ and that the derivative of the left side with respect to x is at most the corresponding derivative of the right side when $x, y \geq 0$. Thus (2.2.13) is true and the proof is completed. □

2.3 Extension to other graphs

This section emphasizes that our results above for the *Pseudo-Tree* model can be extended to other graphs and, in particular, to the analog of models in higher dimensions.

- *Pseudo-tree model in dimension $d' = d + 1$.* Consider the analogous lattice of (2.1.3) in higher dimension

  $$V^d = \{(i, j_1, \ldots, j_d) \in \mathbb{Z}^{d'} : -iC_k \leq j_k \leq iC_k, k = 1, \ldots, d, i \geq 0\}$$

  with oriented edges $(i, j_1, \ldots, j_d) \to (i+1, j_1 + s_1, \ldots, j_d + s_d)$ where $|s_k| \leq C_k \in \mathbb{Z}^+$, $k = 1, \ldots, d$. We denote $\theta_k^d(p)$ to be the probability that there is a significant run of length at least k starting at the origin and p_c^d to be the critical probability. We use the superscript d to emphasize the notation in higher dimension.

With these definitions of the graphs, we have the following results in higher dimension. The proofs of these theorems do not require any argument in addition to what we have already presented, and so they are omitted.

Theorem 2.3.1. *The critical probability of the forgoing pseudo-tree model in dimension $d' = d + 1$ satisfies $p_c^d \geq \frac{1}{(2C_1+1)\times\ldots\times(2C_d+1)}$.*

Theorem 2.3.2. *For $0 < p \leq 1$, there exist positive constants σ_1^d and σ_2^d, independent of p, and there exists a unique function $\phi^d(p)$, which is strictly decreasing and positive when $p < p_c$; constantly 0 otherwise, such that*

$$\sigma_1^d k^{-d} \exp\{-k\phi^d(p)\} \leq \theta_k^d(p) \leq \sigma_2^d k^d \exp\{-k\phi^d(p)\}$$

for any $k \geq 1$. In particular, it follows that

$$-\frac{\log \theta_k^d(p)}{k} \to \phi^d(p). \tag{2.3.14}$$

More generally, let $\mathbb{Z}_+$ be the set of nonnegative integers. For any set $\mathcal{C} \subset \mathbb{Z}_+^d$, we may extend the condition of the oriented edges to a more general condition such as $(i, j_1, \ldots, j_d) \to (i+1, j_1+s_1, \ldots, j_d+s_d)$ where $(s_1, \ldots, s_d) \in \mathcal{C}$. It is straightforward to get the analogous results as above except that $p_c \geq 1/\text{card}\{\mathcal{C}\}$. Details are omitted here.

CHAPTER III

BERNOULLI NET

3.1 Model Introduction

We consider an m-by-n array of nodes, in which there are m rows and n columns. Such an array can be considered as a grid in a two dimensional rectangular region, $([1,n] \times [1,m]) \cap \mathbb{Z}^2$. Assume that each node with coordinate $(i,j), 1 \leq i \leq n, 1 \leq j \leq m$, is associated with a Bernoulli(p) state variable $X_{i,j}$ i.e.,

$$\mathbb{P}(X_{i,j} = 1) = p = 1 - \mathbb{P}(X_{i,j} = 0),$$

where $p \in [0,1]$ is given. Assume state variables of nodes are i.i.d. If $X_{i,j} = 1$, then the node is called significant (or open); otherwise, it is non-significant (or closed). Any two nodes in the grid, say (i_1, j_1) and (i_2, j_2) are *connected* if and only if $|i_1 - i_2| = 1$ and $|j_1 - j_2| \leq C$, with C a prescribed positive integer. Define a chain of length ℓ as a chain of ℓ connected nodes, i.e.,

$$\{(i_1, j_1), (i_1 + 1, j_2), \ldots, (i_1 + \ell - 1, j_\ell) : |j_k - j_{k-1}| \leq C, \forall k = 2, \ldots, \ell\}. \quad (3.1.15)$$

A *significant (or open)* run refers to a chain with all significant nodes. We call such a system *Bernoulli net*. We are interested in the length of the longest significance run in this net. Throughout the thesis, we denote the longest significant run in this net by $L_0(m,n)$ and its length by $|L_0(m,n)|$. Though in some papers *runs*, *chains* and *clusters* have different definitions, here we treat them as synonyms. Such a model is used in the detection of filaments in a point cloud image ([4, 39]) and networks of piecewise polynomial approximation ([25]).

Apparently, the length $|L_0(m,n)|$ depends on parameters n, m, p, and C. Figures 4 and 5 give graphical representations of the relationships between the length $|L_0(m,n)|$

and parameters C, p, m, n. Number of simulations is $1,000$ for each histogram. The following presents a summary of the results.

- For fixed values of m and n, when the value of C or p is increased, the distribution of $|L_0(m,n)|$ changes dramatically. These can be seen in Figure 4.
- For fixed values of C and p, if the value of m or n is doubled, the change of $|L_0(m,n)|$ is not significant. These can be seen in Figure 5.

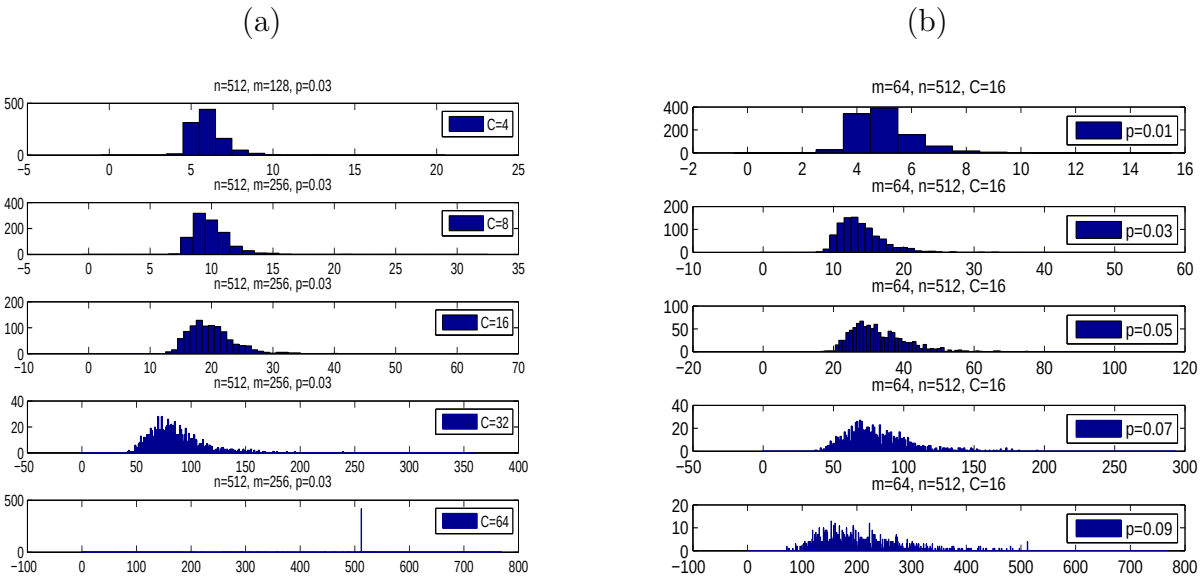

Figure 4: (a) $|L_0(m,n)|$ versus C: effects of connectivity. Every time when the value of C is doubled, the histogram of $|L_0(m,n)|$ is shifted to the right significantly. (b) $|L_0(m,n)|$ versus p: effects of significance probability p. When the value of p is increased, the histogram of $|L_0(m,n)|$ is shifted to the right.

3.2 A thin slab

3.2.1 Previous Work

In this section, we discuss the previous work related to the model in [11]. We will discuss the relationship between $\phi(p)$ mentioned in (2.2.6) and the conditional across probability defined in [11]. We list the results in [11]. For proofs of these results, please refer to [11] and references therein.

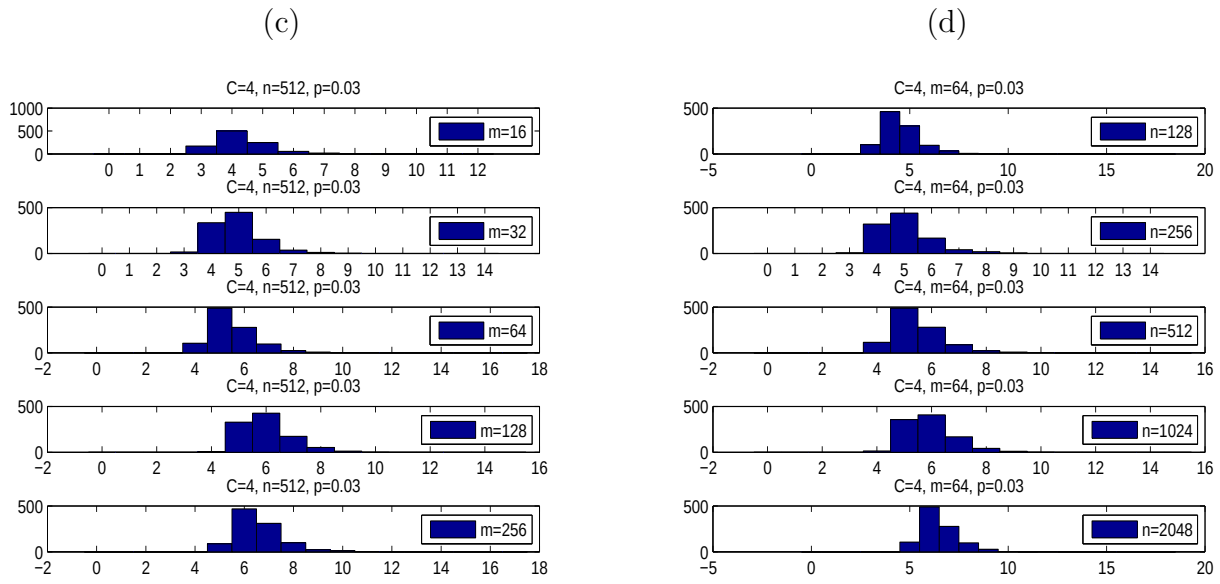

Figure 5: (c) $|L_0(m,n)|$ versus m: effects of heights. When the value of m is doubled, the histogram of $|L_0(m,n)|$ does not change dramatically. (d) $|L_0(m,n)|$ versus n: effects of the width of the Bernoulli net. Every time when the value of n is doubled, the histogram of $|L_0(m,n)|$ does not change dramatically.

The first result is motivated by reliability-focused work [55].

Theorem 3.2.1. *Let $P_k(m,p) = \mathbb{P}_{C,p}(|L_0(m,k)| = k)$ denote the probability that the length of the longest significant run is k, when there are exactly k columns and m rows. We have*

$$(1 - P_k(m,p))^{n-k+1} \leq \mathbb{P}_{C,p}(|L_0(m,n)| < k) \leq [1 - q^m P_k(m,p)]^{n-k+1}, \qquad (3.2.16)$$

where $q = 1 - p$.

The following lemma introduces a constant $\rho(m,p)$ depending on m and p, which is important in the asymptotic distribution of $|L_0(m,n)|$.

Lemma 3.2.2. *Define $\rho_k(m,p) = \frac{P_k(m,p)}{P_{k-1}(m,p)}$. There exists a constant $\rho(m,p)$ in $(0,1)$ that depends on $m, C,$ and p, but not on k such that*

$$\lim_{k\to\infty} \rho_k(m,p) = \rho(m,p).$$

Let an *across* be a significant run that passes all columns from left to right. The ratio $\rho_k(m,p)$ is the conditional probability that conditioning on the fact that there is an across in the previous $(k-1)$ columns, there will be an across for k columns. We may call this the *chance of preserving across significant runs* or *conditional across probability*. The foregoing lemma shows that as the number of columns goes to infinity, the chance of preserving across significant runs converges to a constant.

Now we will recall the result in [11] which is a generalization of the well-known Erdös-Rényi law (See [27, 56, 26]), which is equivalent to the following theorem for $m = 1$ since $\rho(1,p) = p$.

Theorem 3.2.3. *For any fixed $m \in \mathbb{N}$, as $n \to \infty$, we have*

$$\frac{|L_0(m,n)|}{\log_{1/\rho(m,p)} n} \to 1, \quad \textit{almost surely.}$$

Given this theorem, it is easy to obtain the following result which states the relation of ρ and (m,p). Since $|L_0(m,n)|$ actually depends on p, we use the notation $|L_0(m,n,p)|$ in the next corollary to make this dependence explicit.

Corollary 3.2.4. *Given a pair of positive integers m_1, m_2 and a pair of probabilities p_1, p_2 with $m_1 \le m_2$ and $p_1 \le p_2$, we have*

$$\rho(m_1,p_1) \le \rho(m_2,p_1) \quad \textit{and} \quad \rho(m_1,p_1) \le \rho(m_1,p_2)$$

Proof of Corollary 3.2.4. Given a realization

$$t_{i,j} \sim \text{Uniform}(0,1), 1 \le i \le n, 1 \le j \le m,$$

let $x_1^* \ge x_2^* > 0$ be such that $p_1 = \mathbb{P}(t_{i,j} > x_1^*)$ and $p_2 = \mathbb{P}(t_{i,j} > x_2^*)$. Since $t_{i,j} > x_1^*$ implies that $t_{i,j} > x_2^*$, one can easily see that each significant node under threshold x_1^* must be significant under x_2^* and therefore $|L_0(m_1,n,p_1)| \le |L_0(m_1,n,p_2)|$ which by Theorem 3.2.3 leads to $\rho(m_1,p_1) \le \rho(m_1,p_2)$. Similarly, it is not hard to see that $|L_0(m_1,n,p_1)| \le |L_0(m_2,n,p_1)|$ since if $m_1 \le m_2$, $([1,n] \times [1,m_1]) \cap \mathbb{Z}^2 \subset ([1,n] \times [1,m_2]) \cap \mathbb{Z}^2$. Thus $\rho(m_1,p_1) \le \rho(m_2,p_1)$. □

Let us recall the result which states the asymptotic distribution of $|L_0(m,n)|$, the proof of which employs the Chen-Stein approximation method. See [11] and [8].

Theorem 3.2.5. *There exists a constant $A_1 > 0$, that depends only on m, C, and p but not on n, such that for any fixed t, as $n \to \infty$, we have*

$$\mathbb{P}_p(|L_0(m,n)| < \log_{1/\rho(m,p)} n + t) \to \exp\{-A_1 \cdot \rho(m,p)^t\}, \quad as \quad n \to \infty.$$

The analogous result for a one-dimensional Bernoulli sequence is well known. See [29]. The foregoing theorems provide a comprehensive description on the asymptotic distribution of the length of the longest significant run $|L_0(m,n)|$ in a Bernoulli net when the row number m of the array is fixed.

3.2.2 Asymptotic behavior of conditional across probability

We see that all the results in the last subsection depend on $\rho(m,p)$. If $\rho(m,p) \to 1$ as $m \to \infty$, then Theorems 3.2.3 and 3.2.5 may not hold. We shall next discuss the asymptotic behavior of $\rho_k(m,p)$.

Recall that $\theta(p)$ is the probability that there exisits an infinite significant chain rooted at the origin and $p_c = \sup\{p \in [0,1], \theta(p) = 0\}$. We first consider a special case in the array with $m = \infty$ and $n = \infty$. In the following, if $m = \infty$, we employ the lattice of $([1,n] \times \mathbb{Z}) \cap \mathbb{Z}^2$ rather than $([1,n] \times [1,\infty]) \cap \mathbb{Z}^2$. This theorem indicates that as $(m,n) \to (\infty,\infty)$, the behavior of the length of the longest significant run will be quite different in the cases that $p > p_c$ and $p < p_c$.

Theorem 3.2.6. *Let an array have $\mathbb{Z}^+ \times \mathbb{Z}$ nodes, where $\mathbb{Z}^+$ denotes the set of all nonnegative integers. The probability that there exists infinite significant chain (when the marginal probability of a node to be open equal to p), denoted by $\mu(p)$, in the lattice satisfies*

$$\mu(p) = \begin{cases} 0, & \text{if} \quad p < p_c, \\ 1, & \text{if} \quad p > p_c. \end{cases}$$

Proof of Theorem 3.2.6. Recall $\theta(p)$, defined in Subsection 2.2.2, is the probability that there is a significant run starting from a certain node and heading towards right forever when the probability of a node to be open is p. Let $C(i,j)$ be a significant run starting at (i,j). In particular, C is the one starting at $(0,0)$. The event that there exists an infinite open cluster in the array does not depend on the status of finitely many columns of nodes. Thus by the Kolmogorov zero-one law, $\mu(p)$ can only be either 0 or 1. If $p > p_c$, then of course $\theta(p) > 0$. We have

$$\mu(p) \geq \mathbb{P}(|C| = \infty) = \theta(p) > 0,$$

which implies that $\mu(p) = 1$ by the zero-one law. On the other hand, because $\mathbb{P}(|C(i,j)| = \infty) = \mathbb{P}(|C| = \infty) = \theta(p), \forall (i,j)$, if $\theta(p) = 0$ or $p < p_c$, we have

$$\mu(p) \leq \sum_{(i,j)} \mathbb{P}(|C(i,j)| = \infty) = 0.$$

□

We next separate our discussion into super-critical phase $p > p_c$ and sub-critical phase $p < p_c$.

3.2.2.1 Phase $p > p_c$

Our first result shows that in the phase that $p > p_c$, $\rho(m,p) \to 1$ as $m \to \infty$ for any $p > p_c$.

Theorem 3.2.7. *For any $p > p_c$, we have*

$$\lim_{m\to\infty} \rho(m,p) = \rho(\infty,p) = 1 \tag{3.2.17}$$

where $\rho(\infty,p) = \lim_{k\to\infty} \rho_k(\infty,p) = \lim_{k\to\infty} \frac{P_k(\infty,p)}{P_{k-1}(\infty,p)}$, and $\rho_k(\infty,p)$ is the conditional probability that there is an across in the first k columns conditioned on the event that there is an across in the first $k-1$ columns when there are infinitely many rows.

Proof of Theorem 3.2.7. We first prove

$$\rho(\infty, p) = \lim_{k\to\infty} \rho_k(\infty, p) = \lim_{k\to\infty} \frac{P_k(\infty, p)}{P_{k-1}(\infty, p)} = 1,$$

in the case of $p > p_c$. Suppose that $\rho_k(\infty, p)$ does not converge to 1. Then since $[0, 1]$ is a compact set, there must exist a subsequence of $\{\rho_k(\infty, p)\}_{k\in\mathcal{K}}$, $\mathcal{K} \subset \mathbb{Z}^+$ such that

$$\lim_{k(\in\mathcal{K})\to\infty} \rho_k(\infty, p) = \rho_0,$$

for some ρ_0 in $[0, 1)$. And there exists some constant $\rho_0^* \in (0, 1)$ slightly bigger than ρ_0 such that

$$\rho_k(\infty, p) < \rho_0^*,$$

for any sufficiently large $k \in \mathcal{K}$. Therefore, we have

$$P_n(\infty, p) \le \prod_{k(\in\mathcal{K})\le n} \rho_k(\infty, p) \le \prod_{k(\in\mathcal{K})\le n} \rho_0^*, \tag{3.2.18}$$

since $P_n(\infty, p)$, the probability of having an across when there are exactly n columns, is equal to $P_1(\infty, p) \times \prod_{i=1}^{n} \rho_i(\infty, p)$ which is no larger than

$$\prod_{k(\in\mathcal{K})\le n} \rho_0^*.$$

It leads to the fact that

$$P_n(\infty, p) \to 0, \quad \text{as} \quad n \to \infty.$$

On the other hand, it is easy to see that $P_n(\infty, p) \ge \theta_n(p) \ge \theta(p) > 0$ for any n when $p > p_c$, where $\theta_n(p)$ and $\theta(p)$ are defined in subsection 2.2.1. So there is a contradiction. Therefore we should have the following,

$$\rho_k(\infty, p) \to 1, \quad \text{as} \quad k \to \infty,$$

when $p > p_c$.

Now we prove the first equality of (3.2.17) under $p > p_c$. By Corollary 3.2.4, we know the limit of $\rho(m,p)$ exists as m goes to ∞ and thus we have the following

$$\lim_{m\to\infty} \rho(m,p) = \sup_{m\in\mathbb{Z}^+} \{\rho(m,p)\} := \rho^*.$$

If we had $\rho^* < 1$, then notice the fact that $\rho(m,p) \le \rho^*$ for every m, thus it would lead to

$$\frac{|L_0(m,n)|}{\log_{1/\rho(m,p)} n} \to 1, \quad \text{as} \quad n \to \infty, \tag{3.2.19}$$

almost surely, for every m. We would have $|L_0(m,n)| \le \log_{1/\rho^*} n$ with probability 1 when n is sufficiently large for every m. On the other hand, in the array of $\mathbb{Z}^+ \times \mathbb{Z}$, we would have positive probability ($\ge \theta_n(p) \ge \theta(p) > 0$) that there is a significant run connecting the origin and the nth column. This leads to the fact that we have an across in the first n columns with positive probability for any positive integer n. So given n sufficiently large, we may choose $m(\ge 3n\cdot C)$ to be sufficiently large such that the model contains all the possible significant runs in the first n columns starting at the origin. Therefore with positive probability ($> \theta(p)$), we have an across in the first n columns which contradicts (3.2.19) above because $\log_{1/\rho^*} n \ll n$ when n is large. The proof of the theorem is completed. □

We note that $\lim_{m\to\infty} \rho(m,p) = 1$ in the case of $p > p_c$. Recall that we introduce $\phi(p)$ and its property in Corollary 2.2.7. $\phi(p) \equiv 0$ on $p \in [p_c, 1]$. So we have the iterated limit

$$\lim_{m\to\infty} \lim_{k\to\infty} \rho_k(m,p) = \exp\{-\phi(p)\} \tag{3.2.20}$$

when $p \in [p_c, 1]$. Recall that

$$\mathcal{A}_{c_1,c_2,\delta_1,\delta_2} = \{(m,n) : c_1 n^{1+\delta_1} \le m \le c_2 \exp[n(\phi(p) - \delta_2)]\}$$

for positive c_1, c_2, δ_1 and δ_2. In the following, we use $\rho_n(m,p)$ instead of $\rho_k(m,p)$ and we will show below the double limit of $\rho_n(m,p)$ is $\exp\{-\phi(p)\}$ when $p < p_c$ as $n \to \infty$, $m \to \infty$ and $(m,n) \in \mathcal{A}_{c_1,c_2,\delta_1,\delta_2}$ by Chen-Stein's approximation method (See [7]).

3.2.2.2 Phase $p < p_c$

Recall that in Theorem 2.2.4, we introduce $\theta_n(p)$ which is the probability that there is a significant run of size n connecting the origin and $B(n-1)$. In Theorem 3.2.1 we introduce $P_n(m,p)$ which is the probability that the length of the longest significant run is n when there are exactly n columns. To determine the limit of $\rho_n(m,p) = \frac{P_n(m,p)}{P_{n-1}(m,p)}$, we need to know $P_n(m,p)$ when both n and m are very large positive integers.

Theorem 3.2.8. *Let $([1,n]\times[1,m])\cap\mathbb{Z}^2$ be the integer lattice with the probability of nodes being open equal to p. Let $P_n(m,p)$ be the probability of the event that there is a significant run from the first column to the last column of the lattice which is called an across run (or across) in Lemma 3.2.2. Then if $p < p_c$, we have*

$$P_n(m,p) = 1 - \exp\{-m\theta_n(p)\} + o(1),$$

as $m \to \infty, n \to \infty$ and $(m,n) \in \mathcal{A}_{c_1,c_2,\delta_1,\delta_2}$. In particular, we have $\rho_n(m,p) \to \exp\{-\phi(p)\}$ as $m \to \infty, n \to \infty$ and $(m,n) \in \mathcal{A}_{c_1,c_2,\delta_1,\delta_2}$.

Before the proof, let us recall the definitions of three constants in [7]. Let I be an arbitrary index set, and for $\alpha \in I$, let X_α be a Bernoulli random variable with $p_\alpha \equiv \mathbb{P}(X_\alpha = 1) = 1 - \mathbb{P}(X_\alpha = 0) > 0$. For each $\alpha \in I$, suppose we have chose $B_\alpha \subset I$ with $\alpha \in B_\alpha$. We think of B_α as a "neighborhood of dependence" for α, such that X_α is independent or nearly independent of all of the X_β for β not in B_α. Define

$$\begin{aligned} b_1 &\equiv \sum_{\alpha\in I}\sum_{\beta\in B_\alpha} p_\alpha p_\beta, \\ b_2 &\equiv \sum_{\alpha\in I}\sum_{\alpha\neq\beta\in B_\alpha} p_{\alpha\beta}, \text{ where } p_{\alpha\beta} = \mathbb{E}(X_\alpha X_\beta), \\ b_3 &\equiv \sum_{\alpha\in I} s_\alpha, \end{aligned}$$

where

$$s_\alpha \equiv \mathbb{E}\left|\mathbb{E}\{X_\alpha - p_\alpha | \sigma(X_\beta : \beta \in I - B_\alpha)\}\right|.$$

The result in [7] is that when b_1, b_2 and b_3 are all small, then

$$W \equiv \sum_{\alpha \in I} X_\alpha$$

approximately has Poisson distribute with mean

$$\lambda \equiv \mathbb{E}W = \sum_{\alpha \in I} p_\alpha.$$

Proof of Theorem 3.2.8. Let Z_i be the indicator that there is a significant run from $(1, i)$ to the nth column, where $i = 1, \ldots, m$. Let $W_{n,m}$ be the number of nodes in the first column from which an across significant run starts, i.e.,

$$W_{n,m} = \sum_{i=1}^{m} Z_i.$$

Obviously that

$$P_n(m, p) = 1 - \mathbb{P}(W_{n,m} = 0).$$

The main idea of the Poisson approximation is that under certain conditions

$$\mathbb{P}(W_{n,m} = 0)$$

can be approximated by Poisson(λ) where the Poisson parameter λ will be computed below.

To verify the conditions for the Poisson approximation, we first define the neighborhood of i, $1 \leq i \leq m$, as

$$N(i) = \{j : |i - j| < 2 \cdot n \cdot C + 1, 1 \leq j \leq m\}.$$

Define three constants b_1, b_2 and b_3 as in [7] which depend on n, m, C and p. Let $\sigma(Z_j : Z_j \notin N(i))$ be the σ-algebra generated by $\{Z_j : Z_j \notin N(i)\}$. If $j \notin N(i)$, then clearly $|j - i| \geq 2 \cdot n \cdot C + 1$ which leads to the fact that Z_i and Z_j are independent. For b_3, we have

$$\begin{aligned} b_3 &= \sum_{i=1}^{m} \mathbb{E}\left|\mathbb{E}(Z_i - \mathbb{E}(Z_i))\middle|\sigma(Z_j : Z_j \notin N(i))\right| \\ &= 0, \end{aligned}$$

For b_1, we have

$$\begin{aligned} b_1 &= \sum_{i=1}^{m} \sum_{j \in N(i)} \mathbb{E}_p(Z_i)\mathbb{E}_p(Z_j) \\ &= \sum_{i=1}^{m} \sum_{j \in N(i)} \mathbb{P}_p(Z_i = 1)\mathbb{P}_p(Z_j = 1) \\ &\leq \sum_{i=1}^{m} \theta_n(p) \sum_{j \in N(i)} \theta_n(p) \end{aligned}$$

By Theorem 2.2.4, when $p < p_c$, we have a constant $\sigma > 0$ and $\phi(p) > 0$ such that

$$\theta_n(p) \leq \sigma \cdot n \exp\{-n\phi(p)\}.$$

And therefore, it follows that

$$\begin{aligned} b_1 &\leq m \cdot (2n \cdot C + 1) \cdot \sigma^2 \cdot n^2 \cdot \exp\{-2n \cdot \phi(p)\} \\ &\leq O(n^3 \cdot \exp\{-n \cdot (\delta_2 + \phi(p))\}). \end{aligned}$$

For b_2, we have

$$\begin{aligned} b_2 &= \sum_{i=1}^{m} \sum_{j \in N(i), j \neq i} \mathbb{E}_p(Z_i \cdot Z_j) \\ &= 2\sum_{i=1}^{m} \sum_{j \in N(i), j > i} \mathbb{E}_p(Z_i \cdot Z_j) \\ &= 2\sum_{i=1}^{m} \sum_{j \in N(i), j > i} \mathbb{P}_p(Z_i = 1 \quad \text{and} \quad Z_j = 1) \\ &= 2\sum_{i=1}^{m} \mathbb{P}_p(Z_i = 1) \cdot \sum_{j \in N(i), j > i} \mathbb{P}_p(Z_j = 1 | Z_i = 1) \\ &\leq 2\sum_{i=1}^{m} \mathbb{P}_p(Z_i = 1) \cdot (n \cdot C + 1) \\ &\leq O(m \cdot n^2 \cdot \exp\{-n \cdot \phi(p)\}) \\ &\leq O(n^2 \cdot \exp\{-n \cdot \delta_2\}). \end{aligned}$$

In the foregoing, we have used the following fact:

1. $\mathbb{P}(Z_i = 1) \leq \theta_n(p), \forall i = 1, \ldots, m$ and $0 < p < p_c$;

2. $\theta_n(p) \leq O(n \cdot \exp\{-n \cdot \phi(p)\})$;

3. $O(n^{1+\delta_1}) \leq m \leq O(\exp\{n \cdot (\phi(p) - \delta_2)\})$ for some sufficiently small $\delta_1, \delta_2 > 0$.

Loosely speaking, b_1 measures the neighborhood size, b_2 measures the expected number of neighbors of a given occurrence and b_3 measures the dependence between an event and the number of occurrences outside its neighborhood. Now let us consider the Poisson parameter λ which is $\mathbb{E}(W_{n,m})$. When $O(n^{1+\delta_1}) \leq m$ for some sufficiently small $\delta_1 > 0$, by Theorem 2.2.4 it is easy to see that

$$\lambda \approx m\theta_n(p) = m \cdot \exp\{-n \cdot (\phi(p) + o(1))\}$$

as m sufficiently large since $O(n^{1+\delta_1}) \leq m$ is enough to relieve the boundary effects. By Theorem 1 of [7], the Poisson approximation gives

$$\begin{aligned} |\mathbb{P}(W_{n,m} = 0) - \exp\{-\lambda\}| &\leq \min\{1, \frac{1}{\lambda}\} \cdot (b_1 + b_2 + b_3) \\ &\leq O(n^2 \cdot \exp\{-n \cdot (\delta_2 + \phi(p))\}) + O(n^2 \cdot \exp\{-n \cdot \delta_2\}) \\ &\leq O(n^2 \cdot \exp\{-n \cdot \delta_2\}). \end{aligned}$$

Therefore, under the sub-critical phase, i.e., $p < p_c$, if m, n are sufficiently large with $O(n^{1+\delta_1}) \leq m \leq O(\exp\{n \cdot (\phi(p) - \delta_2)\})$, then we have

$$\mathbb{P}(W_{n,m} = 0) = \exp\{-m\theta_n(p)\} + o(1) = \exp\{-m \cdot \exp\{-n \cdot (\phi(p) + o(1))\}\} + o(1).$$

Note that $m\theta_n(p) = m\exp\{-n(\phi(p) + o(1))\} \leq O(\exp\{-n(\delta_2 + o(1)\})$ can be sufficiently small if n is sufficiently large. Since $1 - \exp\{-x\} = x + O(x^2)$ as $x \to 0$, when $p < p_c$ by Corollary 2.2.6 we have

$$\begin{aligned} \rho_n(m,p) &= \frac{P_n(m,p)}{P_{n-1}(m,p)} \\ &= \frac{1 - \exp\{-m\theta_n(p)\} + o(1)}{1 - \exp\{-m\theta_{n-1}(p)\} + o(1)} \\ &= \frac{m\theta_n(p) + O(m^2\theta_n^2(p)\}) + o(1)}{m\theta_{n-1}(p) + O(m^2\theta_{n-1}^2(p)\}) + o(1)} \\ &\to \frac{\theta_n(p)}{\theta_{n-1}(p)} \to \exp\{-\phi(p)\}. \end{aligned}$$

as $m \to \infty, n \to \infty$ and $(m,n) \in \mathcal{A}_{c_1,c_2,\delta_1,\delta_2}$. □

In [11], the authors provide a method to calculates the values of $\rho(m,p)$ (see Table 1) when m is small and fixed by finding out the solution of $\pi = \pi P$ where P is a transition matrix. See also (11) in [11].

Table 1: The values of ρ for different values of m and p, when $C = 1$.

p	0.1	0.2	0.3	0.4	0.5	0.6
m=4	0.2444	0.4564	0.6341	0.7758	0.8804	0.9482
m=8	0.2654	0.4955	0.6869	0.8363	0.9383	0.9876
m=10	0.2691	0.5022	0.6958	0.8467	0.9486	0.9930

One can use simulation to find $\phi(p)$ in the case of $p < p_c$ and thus get some idea about $\rho(m,p)$ as m becomes sufficiently large. See Figure 3. The simulation below is done for the length of the longest significant chain in [11] for $n = 64$, $m = 128$, $C = 3$ and $p = 0.05$ when nodes are assumed to be independent. See Figure 6. The result is based on $10,000$ simulations.

3.3 Rate of the longest significant run

The following is an extension of Theorem 2 in [11] in the case that the Bernoulli net enlarges as $m \to \infty$and $n \to \infty$. In the following, log denotes the logarithm with base e unless the base is explicitly specified.

Theorem 3.3.1. *When $p < p_c$, then as $m \to \infty$ and $n \to \infty$, we have that*

$$\frac{|L_0(m,n)|}{\log(mn)} \to \frac{1}{\phi(p)}, \quad \textit{in probability,} \tag{3.3.21}$$

where $\phi(p)$ is a strictly decreasing, continuous function defined in (2.2.6), which is positive in $(0, p_c)$ and constantly 0 otherwise.

From Theorem 3.3.1, it is apparent that asymptotically m and n do not have a significant impact on the length of the longest significant run $|L_0(m,n)|$. We showed

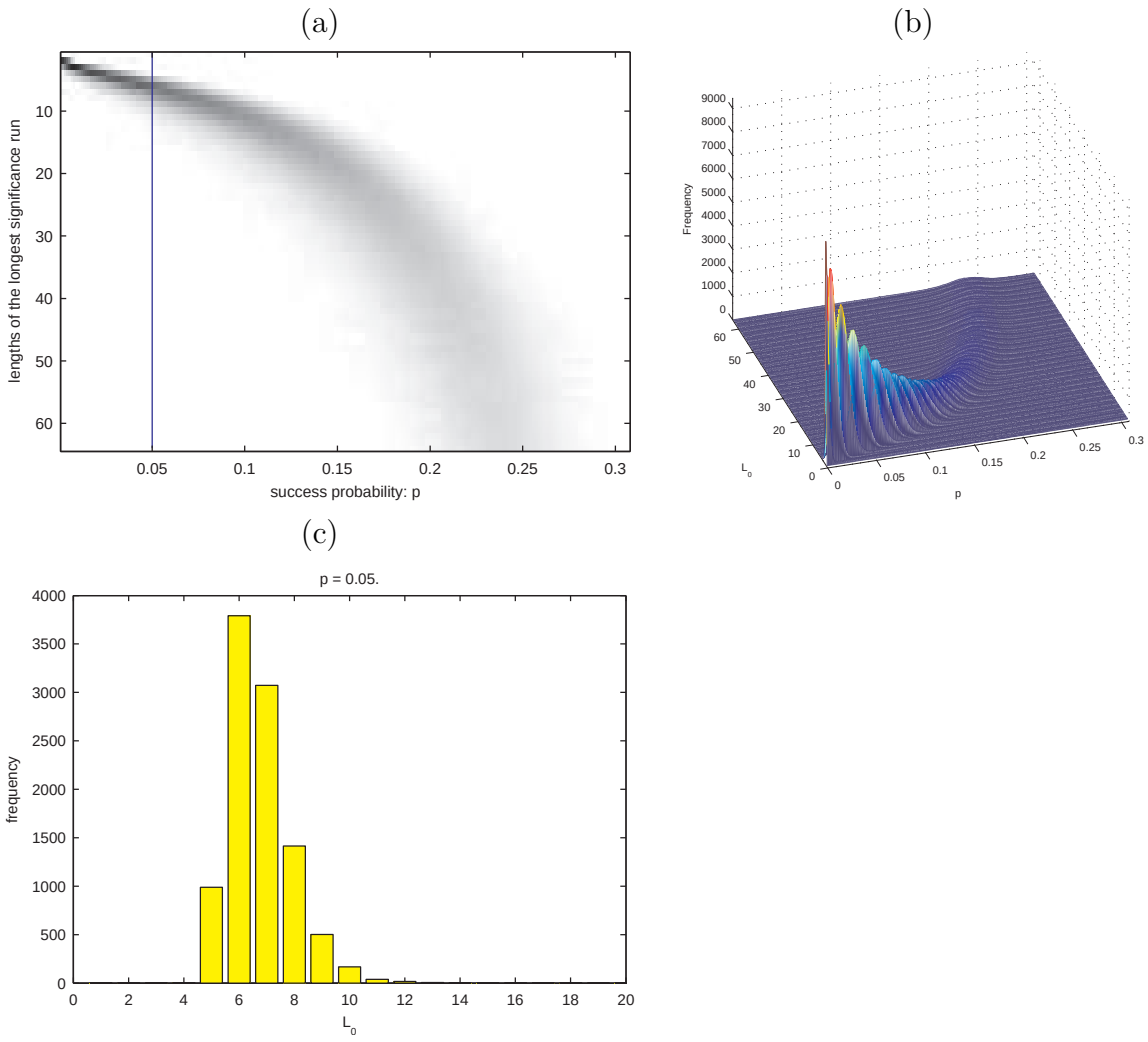

Figure 6: (a) An image plot, the distribution of $|L_0(m,n)|$ (under $n = 64, m = 128, C = 3$) as a function of p ($0 < p < 0.3075$). The intensity of the image is proportional to the frequency of $|L_0(m,n)|$ (which is specified by the y-coordinate) given a value of p (which is the x-coordinate) out of $10,000$ simulations. (b) A mesh plot of the same data as in (a). (c) For $p = 0.05$, the histogram of L_0 based on the same $10,000$ simulations. Note this can be viewed as one vertical slice from (a), or similarly a slice from (b).

that the critical probability $p_c > \frac{1}{2C+1}$ and $|L_0(m,n)|$ will have significantly different asymptotic behaviors between the case $p < p_c$ and $p > p_c$. Therefore, as C and p increases, $|L_0(m,n)|$ will increase dramatically while the increment of m and n do not have a significant impact on the length $|L_0(m,n)|$. Figure 4 and 5 support this argument.

Proof of Theorem 3.3.1. This proof was first used in [6] for a regular lattice model. Recall that $\theta_k^x(p)$ is the probability that $x = (x_1, x_2) \in ([1,m] \times [1,n]) \cap \mathbb{Z}^2$ connects $x_1 + k - 1$-th column, denoted by $B(x_1 + k - 1)$, with a significant chain. One can easily see that $\theta_k^x(p) = \theta_k(p)$. Recall the definition of $\phi(p)$ in the following,

$$\phi(p) = -\lim_{k\to\infty} \frac{\log \theta_k(p)}{k} = -\lim_{k\to\infty} \frac{\log \theta_k^x(p)}{k}.$$

Let $\epsilon < 1/2$ be a small positive number and $k_{m,n}(\epsilon) = \lceil (1+\epsilon)\log(mn)/\phi(p) \rceil$. By the second inequality in (2.2.5), it is not hard to see that

$$\begin{aligned}
\mathbb{P}(|L_0(m,n)| > k_{m,n}(\epsilon)) &= \mathbb{P}\Big(\bigcup_{x\in([1,m]\times[1,n])\cap\mathbb{Z}^2} (x \leftrightarrow B(x_1 + k_{m,n}(\epsilon) - 1))\Big) \\
&\le \sum_{x\in([1,m]\times[1,n])\cap\mathbb{Z}^2} \mathbb{P}(x \leftrightarrow B(x_1 + k_{m,n}(\epsilon) - 1)) \\
&\le mn\sigma_2 k_{m,n}(\epsilon) \exp\{-k_{m,n}(\epsilon)\phi(p)\}
\end{aligned}$$

Since σ_2 is a constant and $\phi(p) > 0$ when $p < p_c$, when m and n are sufficiently large, it follows that

$$\begin{aligned}
mn\sigma_2 k_{m,n}(\epsilon) \exp\{-k_{m,n}(\epsilon)\phi(p)\} &\le mn \exp\{-(1-\epsilon/2)k_{m,n}(\epsilon)\phi(p)\} \\
&\le mn \exp\{-(1-\epsilon/2)(1+\epsilon)\log(mn)\} \\
&= (mn)^{-(\epsilon-\epsilon^2)/2} \\
&\to 0, \quad \text{as} \quad m,n \to \infty
\end{aligned}$$

since $\epsilon - \epsilon^2 > 0$.

On the other hand, let $I = \lceil \frac{mn}{(\log m \log n)^2} \rceil$ and let $k_{m,n}(\epsilon)$ be $\lfloor (1-\epsilon)\log(mn)/\phi(p) \rfloor$. Let $x^1, x^2 \ldots, x^I \in ([1,m] \times [1,n]) \cap \mathbb{Z}^2$ be nodes separated from each other and the

boundary of $([1,m]\times[1,n])\cap\mathbb{Z}^2$ by at least $\frac{1}{2}(\log m\log n)^2$. For sufficiently large m and n, it is not hard to see that the I events $\{x^i \leftrightarrow B(x_1^i + k_{m,n}(\epsilon)-1)\}$ are independent and have equal probabilities. Therefore, for large m and n, by the first inequality of (2.2.5) we have that

$$\begin{aligned}
\mathbb{P}(|L_0(m,n)| < k_{m,n}(\epsilon)) &\leq \mathbb{P}\left(\bigcap_{i=1,\ldots,I}\{\overline{x^i \leftrightarrow B(x_1^i + k_{m,n}(\epsilon)-1)}\}\right)\\
&= \left(1-\mathbb{P}(x^i \leftrightarrow B(x_1^i + k_{m,n}(\epsilon)-1))\right)^I\\
&= (1-\theta_{k_{m,n}(\epsilon)}(p))^I\\
&\leq (1-\sigma_1 k_{m,n}^{-1}(\epsilon)\exp\{-k_{m,n}(\epsilon)\phi(p)\})^I
\end{aligned}$$

When m and n are sufficiently large, it follows that

$$\begin{aligned}
(1-\sigma_1 k_{m,n}^{-1}(\epsilon)\exp\{-k_{m,n}(\epsilon)\phi(p)\})^I &\leq (1-\exp\{-(1+\epsilon/2)k_{m,n}(\epsilon)\phi(p)\})^I\\
&\leq (1-\exp\{-(1+\epsilon/2)(1-\epsilon)\log mn\})^I\\
&\leq (1-(mn)^{-1+\epsilon/2+\epsilon^2/2})^{mn/(\log m\log n)^2}\\
&\leq \left(1-(mn)^{-1+\epsilon/2}\right)^{(mn)^{1-\epsilon/2}(mn)^{\epsilon/2}/(\log m\log n)^2}\\
&\leq \exp\{-(mn)^{\epsilon/2}/(\log m\log n)^2\}\\
&\to 0, \quad \text{as} \quad m,n\to\infty.
\end{aligned}$$

Therefore, as $m,n\to\infty$, we have $\frac{|L_0(m,n)|}{\log mn}\to\frac{1}{\phi(p)}$ in probability. □

3.4 Extension

This section emphasizes that our results above can be extended to the case of models in higher dimensions.

- *Inflating Bernoulli net in dimension $d'=d+1$.* This is the graph with nodes $([1,n]\times[1,m_1]\times\ldots\times[1,m_d])\cap\mathbb{Z}^{d'}$. Assume that each node with coordinate $(i,j_1,\ldots,j_d), 1\leq i\leq n, 1\leq j_k\leq m_k, k=1,\ldots,d$ is associated with a Bernoulli(p) random variables, where $p\in[0,1]$ is given. Equip this

graph with oriented edges $(i, j_1, \ldots, j_d) \to (i+1, j_1 + s_1, \ldots, j_d + s_d)$, where $s_k = 1, \ldots, C_k, k = 1 \ldots, d$ for prescribed $C_k \in \mathbb{Z}^+$. We say a chain to be significant if all the nodes along the chain are significant and denote $L_0(n, m_1, \ldots, m_d)$ to be the longest significant run in this model with length $|L_0(n, m_1, \ldots, m_d)|$.

By Theorem 3.3.1, it is easy to see that we have the following asymptotic rate of the longest significant run.

Theorem 3.4.1. *Let $\phi^d(p)$, defined in (2.3.14), be the higher dimensional version of $\phi(p)$. As $n \to \infty, m_1 \to \infty, \ldots, m_d \to \infty$, we have that*

$$\frac{|L_0(n, m_1, \ldots, m_d)|}{\log(nm_1 \ldots m_d)} \to \frac{1}{\phi^d(p)} \quad \textit{in probability,} \tag{3.4.22}$$

CHAPTER IV

APPLICATIONS

In this chapter, we are going to see some applications of the above theory.

4.1 Detection of an anomalous run in Bernoulli net

In this section, we consider the problem of detecting an anomalous run in Bernoulli net. For simplicity, we only state the low dimension case i.e., $([1,n]\times[1,m])\cap\mathbb{Z}^2$. Let $\mathcal{L}(n,m)$ be a class of chains in $([1,n]\times[1,m])\cap\mathbb{Z}^2$, where a chain is defined as a subset of nodes which is connected as in (3.1.15). Under the null hypothesis, each node (i,j) is i.i.d. associated with random variables $X_{i,j}$ which has Bernoulli distribution of parameter p_0 i.e.,

$$\mathbb{H}_0(n,m): X_{i,j}\sim \text{Bernoulli}(p_0), i.i.d., \forall (i,j).$$

Under the particular alternative where $L\in\mathcal{L}(n,m)$ is anomalous, the variables with index in L have Bernoulli distribution with parameter $p_1>p_0$, i.e.,

$$\mathbb{H}_{1,L}(n,m): X_{i,j}\sim \text{Bernoulli}(p_1), \forall (i,j)\in L; X_{i,j}\sim \text{Bernoulli}(p_0), \forall (i,j)\notin L.$$

Denote the length of the anomalous chain L by $|L|$. For this detection problem, we may consider the test based on the longest significant run in the Bernoulli net $([1,n]\times[1,m])\cap\mathbb{Z}^2$. By Erdös-Rényi law ([26]), the longest significant run in L almost surely has length $\log_{1/p_1}|L|$ as $|L|\to\infty$. Thus if

$$\log_{1/p_1}|L|>\log(nm)/\phi(p), \tag{4.1.23}$$

then the two hypotheses can be separated significantly. Indeed, denote $|L_0(n,m)|$ to be the length of the longest significant run in $([1,n]\times[1,m])\cap\mathbb{Z}^2$ and if

$$|L_0(n,m)|>\log(nm)/\phi(p),$$

then we reject $\mathbb{H}_0(n,m)$; otherwise accept $\mathbb{H}_0(n,m)$. If (4.1.23) holds, then by Theorem 3.3.1, it is easy to see that

$$\mathbb{P}(|L_0(n,m)| > \log(nm)/\phi(p)\big|\mathbb{H}_0(n,m)) + \mathbb{P}(L_0(n,m) \leq \log(nm)/\phi(p)\big|\mathbb{H}_{1,L}(n,m)) \to 0, \tag{4.1.24}$$

as $(n,m) \to (\infty,\infty)$.

For a test T , if $T = 1$, we reject $\mathbb{H}_0$and accept $\mathbb{H}_0$ otherwise; then if

$$\mathbb{P}(T = 0\big|\mathbb{H}_1) + \mathbb{P}(T = 1\big|\mathbb{H}_0) \to 0, \tag{4.1.25}$$

T is called asymptotically powerful test in [5] and this criterion (4.1.25) is widely used in cluster detection literatures (See for example [24, 25, 6, 23]). Thus under the condition (4.1.23), we can see that the test based on the length of the longest significant run is an asymptotically powerful test.

In general, this detection problem can be extended to an exponential model, for instance, the following detection problem in the model with normal distribution,

$$\mathbb{H}_0^N(n,m) : X_{i,j} \sim N(0,1), i.i.d., \forall(i,j);$$

versus given $\mu > 0$

$$\mathbb{H}_{1,L}^N(n,m) : X_{i,j} \sim N(\mu,1), \forall(i,j) \in L; X_{i,j} \sim N(0,1), \forall(i,j) \notin L.$$

After thresholding the values at each node, it is equivalent to the detection problem in the Bernoulli net. We are going to talk about this problem in Chapter 5. The test based on the length of the longest significant chain has also been considered in [6, 46, 45].

4.2 Multi-scale detection of filamentary structure

4.2.1 Background

To be self-contained, we will recall the problem of the length of the longest significant run proposed in [4] in which the authors present a detection method for some

filamentary structure in a background of uniform random points. Suppose we have N data points $X_i \in [0,1]^2$ which at first glance seem to be uniformly distributed in the unit square. Here, for $1 < \alpha \leq 2$, we define that Hölder(α, β) is the class of functions $g : [0,1] \to [0,1]$ with continuous derivative g' that obeys

$$|g'(x) - g'(y)| \leq \alpha\beta \, |x - y|^{\alpha - 1} .$$

Consider the problem of testing

$$\mathbb{H}_0 : X_i \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0,1)^2,$$

versus

$$\mathbb{H}_1(\alpha, \beta) : X_i \overset{\text{i.i.d.}}{\sim} (1 - \epsilon_N)\text{Uniform}(0,1)^2 + \epsilon_N \text{Uniform}(\text{graph}(f)),$$

where $f \in$ Hölder(α, β) is unknown. In words, for the problem of testing, we believe that a relatively small fraction ϵ_n of points lie on a smooth curve in the plane.

In [11], the detection model mentioned in [4] is partially considered and the authors present the convergence rate and the asymptotic distribution of the longest significant run on a Bernoulli Net. However, the row number of the model in [11] is fixed while in [4] the vertical size of the model is increasing very fast when the number of random points tends to infinity. Besides, the nodes in [11] are assumed to be independent while in [4] the nodes are only associated. See [28].

We will review the model in [4] first. Suppose we have N random points uniformly distributed in the square $[0,1] \times [0,1]$. In particular, we use $J = \lceil \log_2(N) \rceil$ to denote its dyadic logarithm. The variable j will index dyadic scales 2^{-j} and will range over $0 \leq j \leq J$. We fix a positive integer $S > 1$ to control the maximum of |slope| we will be able to detect.

Let $R(j, k, \ell_1, \ell_2)$ be a parallelogram with vertical sides that is $\omega = 2^{-j}$ wide by $t = 2^{-(J-j)+1}$ high where j runs through our set of scale indices $\{0, \ldots, J\}$. The

regions in question have a midline that bisects them vertically and will be tilted at a variety of angles. And notice that these regions are highly anisotropic.

The parameters k and $\ell_i, i = 1, 2$, control the horizontal location of the regions and the vertical location and the slope of the midline. There is an underlying assumption that we are only interested in regions whose major axis has a slope bounded in absolute value by S.

To get a vivid impression of this model, see Figure 7 and Figure 8 below. Let $\delta_1 = \frac{t}{4}$ and $\delta_2 = \frac{t}{4\omega}$ (these depend implicitly on j and N). The parallelogram $R(j, k, \ell_1, \ell_2)$ will be centered at $c = ((k + \frac{1}{2})\omega, \ell_1\delta_1)$ and its middle line will have slope $s = \ell_2\delta_2$. Here $0 \le k < \omega^{-1}$, ℓ_1 runs through the set $0, \ldots, \delta_1^{-1} - 1$ and ℓ_2 runs through the set $-S\delta_2^{-1}, \ldots, 0, \ldots, S\delta_2^{-1}$. We gather all such regions at level (scale) j in $\mathcal{R}(j) = \{R(j, k, \ell_1, \ell_2) : k, \ell_1, \ell_2\}$ and therefore we have $2^j \times 2^{J-j+1} \times S \cdot 2^{J-2j+2} + 1$ or $O(N^2)$ parallelograms in total. To organize the regions, we define a directed graph $\mathcal{G}(j) = (\mathcal{V}(j), \mathcal{E}(j))$, with vertices $\mathcal{V}(j)$ and edges $\mathcal{E}(j)$.

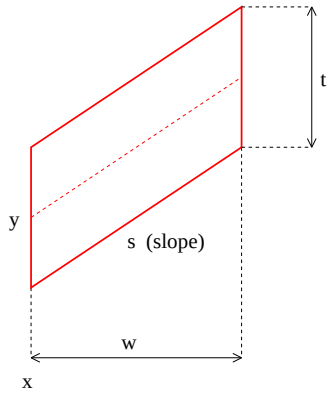

Figure 7: An Anisotropic 'Strip' R

The vertices are simply the regions $\mathcal{R}(j)$, i.e., $\mathcal{V}(j) \equiv \mathcal{R}(j)$. The edges connect regions by good continuation, namely, to regions that are horizontally adjacent, and that have altitudes and slopes that are nearly the same-less than t and $\frac{t}{\omega}$ apart, respectively. Formally, we have the directed edges in $\mathcal{E}(j)$ as

$$(k, \ell_1, \ell_2) \to (k + 1, \ell_1 + \ell_2 + u, \ell_2 + v), \tag{4.2.26}$$

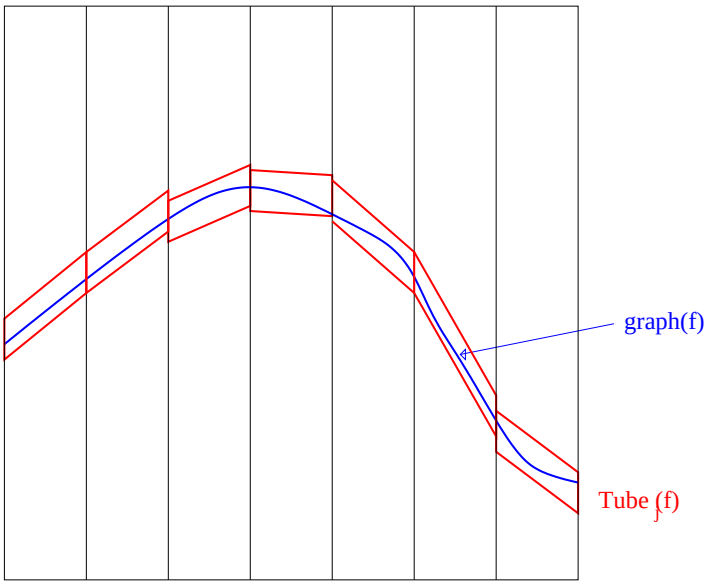

Figure 8: $graph(f)$ (in blue) covered by $Tube_j(f)$ (in red).

where $|u| \leq 4, |v| \leq 4$ and we call (4.2.26) the connectivity of edges. The mapping between these discrete parameters is intended to insure that the regions pack together horizontally and that they are fairly closely spaced in both vertical position and slope.

For every region $R \in \mathcal{R}(j)$, we count the number of the points that fall into R, denoted by $N(R, j)$. We define a significance indicator, which is nonzero when the counts $N(R, j)$ exceeds a prescribed threshold N^*, i.e.,

$$s(R) = \mathbf{1}_{\{N(R,j)>N^*\}}. \tag{4.2.27}$$

We say that N^* is the counting threshold in the following. The significance indicator may be viewed as a label on the regions R, producing a sequence of a labeled graphs

$$\Sigma(j) = (\mathcal{V}(j), \mathcal{E}(j), \sigma(j)),$$

where $\sigma(j) = (s(R))$ gives the labels on $R \in \mathcal{R}(j)$. We call this the j-th significance graph.

In each significance graph, we employ a depth-first search algorithm to explore all significance paths

$$\pi = (R_1, R_2, \ldots, R_m),$$

that is, sequence of vertices that are:

(a) all significant, $s(R_k) = 1$;

(b) all connected, $(R_k, R_{k+1}) \in \mathcal{E}(j)$.

We record the maximum path length in each significant graph as follows:

$$\left|L_{N,j}^{\max}\right| = \max\{\text{length}(\pi) : \pi \quad \text{is a significant path in} \quad \Sigma(j)\},$$

$$|L_N^{\max}| = \max_j \left|L_{N,j}^{\max}\right|.$$

The decision of the hypothesis testing problem is that we compare $|L_N^{\max}|$ with a length threshold: If $|L_N^{\max}| \leq |L_N^*|$, accept $\mathbb{H}_0$; if $|L_N^{\max}| > |L_N^*|$, then reject $\mathbb{H}_0$.

We call $|L_N^*|$ the decision threshold in the following. Under the assumption that N points are randomly distributed in the square $[0,1] \times [0,1]$, the counting threshold determines the probability of $\{s(R) = 1\}$. Because the area of each region is $\frac{2}{N}$, we have the following,

$$\mathbb{P}(s(R) = 1) = \mathbb{P}(\text{Bin}(N, \frac{2}{N}) > N^*),$$

where $\text{Bin}(N, \frac{2}{N})$ denotes the random variable with Binomial distribution of parameters N and $\frac{2}{N}$. Because Poisson(2) is an approximation of $\text{Bin}(N, \frac{2}{N})$ when N is sufficiently large, we use Poisson(2) instead in the following. The main result of this multi-scale detection method in [4] is the following:

Theorem 4.2.1. *There is a single choice of threshold N^* and $|L_N^*|$ so that for every $\alpha \in (1,2]$ and $\beta > 0$, there is $T_*(\alpha, \beta, S)$ such that for each $\epsilon_N > T_* N^{-\alpha/(1+\alpha)}$*

$$\mathbb{P}(\textit{test rejects } \mathbb{H}_0 \big| \mathbb{H}_1(\alpha, \beta, S)) \to 1, \quad \textit{as} \quad n \to \infty,$$

and at the same time

$$\mathbb{P}(\textit{test rejects } \mathbb{H}_0 \big| \mathbb{H}_0) \to 0, \quad \textit{as} \quad n \to \infty.$$

We give the specifications of the foregoing thresholds. In [4], the authors define N^* such that

$$p = \mathbb{P}(\text{Poisson}(2) > N^*) < \frac{p_0}{81}, \tag{4.2.28}$$

where $p_0 \in (0,1)$ is some chosen number. With the help of Erdös-R*ényi* law, the authors define the decision threshold

$$|L_N^*| \equiv 3 \log_{1/p_0} N. \tag{4.2.29}$$

The specification of $T_*(\alpha, \beta, S)$ in Theorem 4.2.1 is a little bit complicated. First define

$$p^* = p_0^{\frac{1}{18}}, \tag{4.2.30}$$

Then let λ^* be a constant that satisfies

$$\mathbb{P}(\text{Poisson}(\lambda^*) < N^*) \leq \frac{1 - p*}{2}.$$

Then $T_*(\alpha, \beta, S) = 2\lambda^* \beta^{\frac{1}{1+\alpha}} \sqrt{1 + S^2}$. See [4] for details.

4.2.2 A revisit using the theory of longest chain

In this part, we will apply our theory to the model in [4] for the detection problem. Consider the problem of testing

$$\mathbb{H}_0 : X_i \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0,1)^2,$$

versus

$$\mathbb{H}_1(\alpha, \beta) : X_i \overset{\text{i.i.d.}}{\sim} (1 - \epsilon_N)\text{Uniform}(0,1)^2 + \epsilon_N \text{Uniform}(\text{graph}(f)),$$

where $f \in \text{Hölder}(\alpha, \beta)$ is unknown, N is the total number of points in $[0,1] \times [0,1]$ and $\epsilon_N > T_* N^{-\frac{\alpha}{1+\alpha}}$ is the portion of the points lying on the graph of the function. We can see that when the number of random points N in $[0,1] \times [0,1]$ tends to infinity, the background of uniform random points can be treated as sampled from a (spatial) Poisson process. One of the properties of this Poisson process is that for

any subregion Ω in the unit square, the number of points in this region, denoted by $N(\Omega)$, also has the Poisson distribution with parameter $N \cdot |\Omega|$, where $|\Omega|$ is the area of Ω, i.e., $N(\Omega) \sim \text{Poi}(N \cdot |\Omega|)$. Another property of the (spatial) Poisson process is that for any two non-overlapping regions Ω_1 and Ω_2 in the unit square, the number of points in Ω_1 and the number of points in Ω_2, i.e., $N(\Omega_1)$ and $N(\Omega_2)$ respectively, are independent.

Let us now rephrase the main results of [28] here.

Definition 4.2.2. *Let $T_1, T_2, \ldots, T_n$ be n associated random variables if $Cov[f(\boldsymbol{T}), g(\boldsymbol{T})] \geq 0$, where $\boldsymbol{T} = (T_1, T_2, \ldots, T_n)$, for all nondecreasing functions f and g, for which the expectations $\mathbb{E}(f), \mathbb{E}(g)$ and $\mathbb{E}(fg)$ exist.*

It is known that

- any subset of associated random variables are associated;
- nondecreasing functions of associated random variables are associated;
- independent random variables are associated
- let $x_1, \ldots, x_n$ be associated binary random variables, then

$$\mathbb{P}(x_1 = s, \ldots, x_n = s) \geq \mathbb{P}(x_1 = s) \ldots \mathbb{P}(x_n = s),$$

  where s can be either 0 or 1.

In the method of [4], as the number of points in the unit square tends to infinity, one can easily see that the number of random points in two parallelograms R_1 and R_2 are independent if R_1 and R_2 are non-overlapping, and they are associated if R_1 and R_2 are overlapping. Indeed, if we use $s(R)$ to denote the state (significant or non-significant) of parallelogram R, then we have

$$\mathbb{P}(s(R_1) = a, s(R_2) = a) \geq \mathbb{P}(s(R_1) = a)\mathbb{P}(s(R_2) = a), \tag{4.2.31}$$

where a is either 0 or 1. The equality in (4.2.31) holds when R_1 does not overlap with R_2.

For multi-scale detection problem, we construct an array of nodes in

$$\mathcal{V} \equiv [1, 2^j] \times [1, 2^{J-j+1}] \times [-S2^{J-2j+1}, S2^{J-2j+1}] \cap \mathbb{Z}^3, \tag{4.2.32}$$

where $J = \lceil \log_2(N) \rceil$ and $0 \leq j \leq J$. For any nodes

$$(k, \ell_1, \ell_2) \in [1, 2^j] \times [1, 2^{J-j+1}] \times [-S2^{J-2j+1}, S2^{J-2j+1}] \cap \mathbb{Z}^3$$

in the array, the three components represent the location index, the altitude index and the slope index, respectively. In light of the nodes in two dimension, we might consider $m = 2^{J-j+1} \times (2S \cdot 2^{J-2j+1} + 1)$ nodes in the same strip as a column and thus there are $n = 2^j$ columns in total. For any node (k, ℓ_1, ℓ_2), it can be connected to

$$(k+1, \ell_1 + \ell_2 + u, \ell_2 + v) \in [1, 2^j] \times [1, 2^{J-j+1}] \times [-S2^{J-2j+1}, S2^{J-2j+1}] \cap \mathbb{Z}^3, \tag{4.2.33}$$

where $|u| \leq 4$, $|v| \leq 4$. Each node is associated with a parallelogram in the algorithm mentioned in [4] and therefore it is open with probability

$$p = \mathbb{P}(N(R) > N^*) \to \mathbb{P}(\text{Poisson}(2) > N^*), \quad \text{as} \quad N \to \infty$$

where $N(R)$ is the number of points in the parallelogram R and N^* is a counting threshold to be specified later. Due to the structure of the model in [4], the nodes in different columns are independent and all the nodes here are associated as $N \to \infty$.

Consider the Pseudo-tree model in dimension 3, as in Section 2.3,

$$V^2 = \{(i, j_1, j_2) \in \mathbb{Z}^2 : -4i \leq j_1 \leq 4i, -4i \leq j_2 \leq 4i, i \geq 0\},$$

with oriented edges $(i, j_1, j_2) \to (i+1, j_1 + s_1, j_2 + s_2)$, where $|s_1| \leq 4$ and $|s_2| \leq 4$. We denote $\theta_k^2(p)$ to be the probability that there is a significant run of length at least k starting at the origin and p_c^2 to be the critical probability. Revisiting the proofs of Theorems 2.2.1, 2.2.4 and 3.3.1 together with their generalized results in Theorems

2.3.1, 2.3.2 and 3.4.1, we find that these results do not depend on the independence of nodes in the same column. The condition that nodes are associated in the same strip is sufficient for these theorems. By Theorem 2.3.2, there exist positive constants σ_1^2 and σ_d^2, independent of p, and there exists a unique function $\phi^2(p)$, which is strictly decreasing and positive when $p < p_c^2$; constantly 0 otherwise, such that

$$\sigma_1^2 k^{-2} \exp\{-k\phi^2(p)\} \leq \theta_k^2(p) \leq \sigma_2^2 k^2 \exp\{-k\phi^2(p)\}$$

for any $k \geq 1$. In particular, it follows that

$$-\frac{\log \theta_k^2(p)}{k} \to \phi^2(p).$$

Since each node in the array can be connected with at most 81 nodes in the next column and hence $p_c^2 \geq \frac{1}{81}$ by Theorem 2.3.1.

Though in [4] the authors consider all scales in $\{j : 0 \leq j \leq J \text{ for } J = \lceil \log_2 N \rceil\}$, we will consider $\{j : 0 \leq j \leq \lceil \frac{J+\log_2 \beta}{1+\alpha} \rceil\}$ only. We shall point out here that the restriction on j is a fairly reasonable assumption for the following reasons. First notice that if we choose $j > \lceil \frac{J+\log_2 \beta}{1+\alpha} \rceil$, then the range of the slope index $[-S2^{J-2j+1}, S2^{J-2j+1}]$ will be fairly small. Hence the parallelograms will be almost horizontal rectangles. Moreover, under $\mathbb{H}_1(\alpha, \beta)$, for scales $j \leq \lceil \frac{J+\log_2 \beta}{1+\alpha} \rceil$, the parallelograms in the same column will be more overlapping which yields more significant nodes and hence the longer length of the significant runs. And it is easier to separate the null hypothesis $\mathbb{H}_0$ from the alternative hypotheses $\mathbb{H}_1(\alpha, \beta)$. The most important reason is that, in [4], the authors point out that under $\mathbb{H}_1(\alpha, \beta)$, there is some scalar j^* such that the graph of the function is completely covered by a tube of parallelograms in this scale like the case in Figure 8. We call this containing tube $T_{j^*}(f)$. It is shown that $j^* = \lceil \frac{J+\log_2 \beta}{\alpha+1} \rceil$ (See Lemma 2.1-2.3 and their proofs in [4]). In other words, using only scalars $j \leq \lceil \frac{J+\log_2 \beta}{1+\alpha} \rceil$ is enough to cover the graph hence detect the filamentary structure under $\mathbb{H}_1(\alpha, \beta)$. Thus, it actually can save work to consider only the scales no larger than $\lceil \frac{J+\log_2 \beta}{1+\alpha} \rceil$ without loss of generality. In case that $\alpha \in (1, 2]$ and $\beta > 0$

are unknown, it is possible to use $0.5001J$ instead of $\lceil \frac{J+\log_2 \beta}{\alpha+1} \rceil$ for the reason that $\lceil \frac{J+\log_2 \beta}{\alpha+1} \rceil \leq 0.5001J$ as $J \to \infty$. Denote $\lceil \frac{J+\log_2 \beta}{\alpha+1} \rceil$ by c_J, which is the scale under which the whole graph of the function is guaranteed to be in a series of parallelograms, as shown in Figure 8.

Now we specify the asymptotic thresholds for our purpose. These thresholds are better and more intuitive than those in [4]. Let the membership threshold N^*, as in (4.2.28), satisfies the following property:

$$p_0 = \mathbb{P}(\text{Poisson}(2) > N^*) < \frac{1}{81} \leq p_c^2.$$

Here we can take $N^* = 6$ so that $p_0 \approx 0.0045338 < \frac{1}{81}$. Let the decision threshold $|L_N^*|$, as in (4.2.29), be

$$(1+\delta_3)\frac{2J\log 2}{\phi(p_0)},$$

for some small $\delta_3 > 0$.

Define p^*, as in (4.2.30), to be

$$\exp\{-\phi(p_0)\frac{c_J(1-\delta_3)}{2J(1+\delta_3)}\} \tag{4.2.34}$$

We choose λ^* such that

$$\mathbb{P}(\text{Poisson}(\lambda^*) > N^*) > p^*.$$

Finally, we define $T_*(\alpha, \beta, S)$, as in Theorem 4.2.1, to be $2\lambda^* \beta^{\frac{1}{1+\alpha}} \sqrt{1+S^2}$. It is shown in (4.5) of [4] that if $\epsilon_N > T_*(\alpha, \beta, S) N^{-\frac{\alpha}{1+\alpha}}$, then for each parallelogram R in the containing tube $T_{j^*}(f)$ like the case in Figure 8, we have

$$\mathbb{P}(N(R) > N^*) > p^*, \tag{4.2.35}$$

where $N(R)$ is the number of points in the parallelogram R.

Let $|L_N^{\max}|$ denote the length of the longest significant run in Lattice $\mathcal{V}$, defined in (4.2.32). Note that there are $4S2^{2J-2j} + 2^{J+1}$ nodes in $\mathcal{V}$. Under $\mathbb{H}_0$, by Theorem 3.4.1, for any small ϵ and $\delta_3 > 0$, with probability at least $1-\epsilon$, we have

$$|L_N^{\max}| \leq (1+\delta_3/2)\frac{\log(4S2^{2J-2j}+2^{J+1})}{\phi(p_0)} \leq (1+\delta_3)\frac{2J\log 2}{\phi(p_0)}, \tag{4.2.36}$$

as $N \to \infty$. Similarly, under $\mathbb{H}_1(\alpha, \beta)$, with probability at least $1-\epsilon$ and for large N, by the Erdös- Rényi Law and the Egoroff's Theorem ([63]), we have that the length of the significant run in the tube $T_{j^*}(f)$ containing the function f, denoted by $|L_{j^*}(f)|$, satisfies

$$\begin{aligned} |L_{j^*}(f)| &> (1-\delta_3) c_J \log_{1/p^*} 2 && (4.2.37) \\ &= (1+\delta_3) \frac{2J \log 2}{\phi(p_0)} && (4.2.38) \\ &= |L_N^{\max}|, && (4.2.39) \end{aligned}$$

as $N \to \infty$. The inequality (4.2.37) is due to the fact that (4.2.35) holds for each parallelogram in the containing tube $T_{j^*}(f)$. The equality (4.2.38) is due to the definition of p^* in (4.2.34). Therefore when $\epsilon_N > T_+ N^{-\frac{\alpha}{1+\alpha}}$, by (4.2.36) to (4.2.39), we have asymptotically powerful test based on the length of the longest significant chain, i.e.,

$$\begin{aligned} \mathbb{P}(|L_N^{\max}| > |L_N^*| \,\big|\mathbb{H}_0) &\to 0, \quad \text{as} \quad N \to \infty, && (4.2.40) \\ \mathbb{P}(|L_N^{\max}| < |L_N^*| \,\big|\mathbb{H}_1(\alpha, \beta)) &\to 0, \quad \text{as} \quad N \to \infty. && (4.2.41) \end{aligned}$$

4.3 Target tracking problems

4.3.1 Background

In this subsection, we discuss another application of the theory. Let $X_i \in \{0,1\}^m$, where m is an integer. We have $X_{i,j} = 0$ or 1, where $X_{i,j}$ denotes the jth entry of X_i. Here i is a time index and j is a location index. $X_{i,j} = 1$ (or 0) corresponds to a target being present (or absent) at location j and time i.

We introduce the following probabilistic model to minic the motion of targets over time. From X_i to $X_{i+1}, 1 \le i \le n-1$, we have:

1. Initialize $X_{i+1,j} = 0$ for all j.

2. If $X_{i,j} = 0$, then set $X_{i+1,j} = X_{i+1,j} + 1$ with probability p_0 (corresponding to a newly emerging object).

3. If $X_{i,j} = 1$, there are four sub-cases:

   (a) $X_{i+1,j-1} = X_{i+1,j-1} + 1$ with probability p_1 (shifting left)

   (b) $X_{i+1,j} = X_{i+1,j} + 1$ with probability p_2 (remain the same location)

   (c) $X_{i+1,j+1} = X_{i+1,j+1} + 1$ with probability p_3 (shifting right)

   (d) do nothing, with probability $1 - p_1 - p_2 - p_3$ (object vanishes).

   Apparently, we must impose $0 < p_i < 1$, for $i = 0, 1, 2, 3$ and $p_1 + p_2 + p_3 < 1$.

4. Finally, we take $X_{i+1,j} = \min(1, X_{i+1,j})$ to ensure that each one of them is either one or zero. Note X form the ground truth regarding the presence and locations of the targets.

   Below we consider how observations are generated.

5. Set $z_{ij} = x_{ij} + \epsilon_{ij}$, where $\epsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, where σ^2 is a parameter, e.g., $\sigma^2 = 1$.

   Note $Z_i = \{z_{ij}, j = 1, 2, \ldots, m\}$ is the observation at time i.

In [59], a hidden state Markov process model is mentioned. In the above case, it is as follows:

$$\begin{array}{ccccccccccc} X_0 & \to & X_1 & \to & X_2 & \to & X_3 & \to & X_4 & \to \cdots \\ & & \downarrow & & \downarrow & & \downarrow & & \downarrow & & \\ & & Z_1 & & Z_2 & & Z_3 & & Z_4 & . & \end{array}$$

For our purpose, we may not emphasize this Markovian aspect of the problem. It is important in the estimation.

We pose a hypothesis testing problem in this case, i.e.,

$$\mathbb{H}_0: \quad \text{all} \quad X_{i,j} = 0 \quad \text{versus} \quad \mathbb{H}_1: \quad \text{some} \quad X_{i,j} = 1.$$

The idea behind the hypothesis testing problem is to say whether there is some newly emerging object at certain location and time or the image just consists of white noisy pixels. We will use the theory of the longest chain to solve this problem in the following subsection.

4.3.2 A revisit using the theory of longest chain

In this part, we will use our theory to estimate an upper bound of the length of the longest significant run in the target tracking problem in an array of size m-by-n. Under the null hypothesis $\mathbb{H}_0$, the pixel image of size $[1,m]\times[1,n]$ is just a white noise image and $Z_{i,j}=\epsilon_{i,j}$ where $\epsilon_{i,j}\overset{\text{i.i.d.}}{\sim}N(0,\sigma^2)$. For an arbitrary node $Z_{i,j}$ to be significant, we should provide a member threshold Z^*, i.e., the node is significant if $Z_{i,j}>Z^*$ and insignificant otherwise. Let $\mathcal{V}$ be the set of nodes under consideration, i.e.,

$$\mathcal{V}\equiv\{(i,j)\in\mathbb{Z}^2:1\le i\le n,1\le j\le m\}.$$

Let $\mathcal{E}$ be the set of edges from $(i,j)\in\mathcal{V}$ to $(i+1,j+s)\in\mathcal{V}$ such that $|s|\le 1$. Let $p=\mathbb{P}(Z_{i,j}>Z^*)$ be the probability of a node $(i,j)\in\mathcal{V}$ to be significant. In order to make a decision, we need to count the length of the longest significant nodes among all the chains along the edges in $\mathcal{E}$, i.e., the chains of the following form

$$\{(i,j_0),(i+1,j_1),\dots,(i+\ell,j_\ell):|j_{k+1}-j_k|\le 1,k=0,\dots,\ell-1\}.$$

We will use the length of the longest significant run, denoted by $|L_T^{\max}(m,n)|$, as a statistic for the test. And a little bit more consideration yields that under the null hypothesis $\mathbb{H}_0$, the chain of significant nodes has the same structure as in (3.1.15) with $C=1$.

We can apply our theory to find a reasonable threshold. By Theorem 2.2.1, the critical probability p_c for the graph $(\mathcal{V},\mathcal{E})$ satisfies that $p_c\ge\frac{1}{2C+1}$. Therefore, we may choose Z^* such that $p=\mathbb{P}(Z_{i,j}>Z^*)<\frac{1}{3}$ for $(i,j)\in\mathcal{V}$. Thus if m is constant, by Theorem 3.2.3, for any $\epsilon_1>0$, there exist $\rho(m,p)\in(0,1)$ and $N\in\mathbb{Z}^+$ such that when $n\ge N$, we have

$$|L_T^{\max}(m,n)|\le(1+\delta_4)\log_{1/\rho(m,p)}n \text{ with probability } 1-\epsilon_1,$$

for any $\delta_4 > 0$. If $m \to \infty$, $n \to \infty$, then by Theorem 3.3.1, for any $\epsilon_2 > 0$ we have

$$|L_T^{\max}(m,n)| \leq (1+\delta_4)\frac{\log(mn)}{\phi(p)} \text{ with probability } 1-\epsilon_2.$$

So let the decision threshold $|L_T^*|$ be $(1+\delta_4)\log_{1/\rho(m,p)} n$ if $n \to \infty$ with fixed m; and $(1+\delta_4)\frac{\log(mn)}{\phi(p)}$ if $n \to \infty, m \to \infty$. Since the forgoing ϵ_1 and ϵ_2 are arbitray, if it happens that

$$|L_T^{\max}(m,n)| > |L_T^*|,$$

then we can always reject $\mathbb{H}_0$ with false positive close to 0 asymptotically.

CHAPTER V

FAST AND NEAR-OPTIMAL ALGORITHMS TO DETECT FILAMENTARY OBJECTS IN DIGITAL IMAGES

5.1 Statistical model

We consider an m-by-n array of nodes $\mathcal{S}$ with m rows and n columns, i.e.,

$$\mathcal{S} = \{(i, j) : 1 \leq i \leq n, 1 \leq j \leq m\}. \tag{5.1.42}$$

We assume $m \in \mathbb{Z}^+$ is fixed and will eliminate this restriction in Section 5.8. Such an array can be considered as a grid in two dimensional rectangular region $[1, n] \times [1, m]$. Assume that each node with coordinate $(i, j) \in \mathcal{S}$ is associated with a normal distribution $X_{i,j}$. In a digital image, each node indicates a pixel of the image and its corresponding normal random variable $X_{i,j}$ denotes the intensity of the image at (i, j). We consider all the embedded chains in $\mathcal{S}$ with good continuation such that nodes on the chain are horizontally adjacent and their altitudes are nearly the same—less than $C \in \mathbb{Z}^+$ apart for neighboring nodes. To be precise, for $(i, j_0) \in \mathcal{S}$ and $L \geq 0$, a chain $\mathcal{L}$ of good continuation with length $L + 1$ has the following form,

$$\mathcal{L} = \{(i, j_0), (i + 1, j_1), \ldots, (i + L, j_L) : |j_k - j_{k-1}| \leq C, 1 \leq k \leq L\}. \tag{5.1.43}$$

As a first attempt to formalize matters, consider the problem of testing

$$\mathbb{H}_0 : X_{i,j} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2), \forall (i, j) \in \mathcal{S} \tag{5.1.44}$$

versus

$$\begin{aligned} \mathbb{H}_1(\mathcal{L}_n^0, \mu) : \quad & X_{i,j} \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2), \text{ for some } \mu > 0, \text{ when } (i, j) \in \mathcal{L}_n^0, \\ & X_{i,j} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2), \forall (i, j) \in \mathcal{S} \setminus \mathcal{L}_n^0, \end{aligned} \tag{5.1.45}$$

where $N(\mu, \sigma^2)$ stands for a normal distribution with mean μ and standard deviation σ and $\mathcal{L}_n^0$ is a chain in $\mathcal{S}$ with good continuation. We explicitly indicate that $\mathcal{L}_n^0$ depends on n because except Section 5.8, we assume that m is fixed and we want to handle the detectability of (5.1.44) versus (5.1.45) asymptotically as the number of columns $n \to \infty$. For convenience, in this paper, we assume $\sigma = 1$. Note that by varying the location and orientation of embedded chain $\mathcal{L}_n^0$ and the value of the parameter μ, there are infinite number of possibilities under the alternative hypothesis $H_1(\mathcal{L}_n^0, \mu)$. The objective of our forgoing testing problem is to detect whether there is an embedded chain $\mathcal{L}_n^0$ in $\mathcal{S}$ with an elevated mean $\mu > 0$ in the digital image $\mathcal{S}$. More specifically, how large the value of μ and how long the chain $\mathcal{L}_n^0$ should be so that the corresponding alternative hypothesis can be strongly distinguished from the null hypothesis. Throughout this section, the length of the chain and the number of nodes on the chain are the same. We use $|\cdot|$ to indicate the length of a chain or the cardinality of a set. We use $C, C_1, C_2, \delta_1, \delta_2, \eta, \eta_1, \eta_2, \zeta$ to indicate positive constants which may vary line by line.

5.2 Related work and existing results

In [5], the authors consider the following problem detecting intervals in dimension one. Let $X = (x(i) : 0 \le i < n)$ be an array of random variables which contain white noise, except possibly on an interval where the mean might be elevated, i.e.,

$$x(i) = \mu_n 1_{\{a \le i < b\}} + z(i), i = 0, \ldots, n-1.$$

Here, the endpoints a, b of the interval obey $0 \le a < b \le n$ but are assumed to be unknown; and μ_n is the amplitude of the signal and $z(i)$ are i.i.d. standard normal random variables. Obviously, one can see that this problem is a special case of our problem when $m = 1$. Let $A_n = \mu_n \times \sqrt{b' - a'}$ and $\xi_{a',b'}(i) = 1_{\{a' \le i < b'\}}/\sqrt{b' - a'}$ be the normalized prototype of an interval, and let

$$X_n^* = \max_{0 \le a' < b' \le n} \langle \xi_{a',b'}, X \rangle, \tag{5.2.46}$$

where $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is the inner product.

For completeness, we give the following definition of an asymptotically powerful statistic and asymptotically powerless testing problems in [5].

Definition 5.2.1. *In a sequence of testing problems* $(\mathbb{H}_{0,n})$ *versus* $(\mathbb{H}_{1,n})$*, we say that a sequence of tests* T_n *is asymptotically powerful if*

$$\mathbb{P}_{\mathbb{H}_{0,n}}\{T_n \text{ rejects } \mathbb{H}_{0,n}\} + \mathbb{P}_{\mathbb{H}_{1,n}}\{T_n \text{ accepts } \mathbb{H}_{0,n}\} \to 0,$$

as $n \to \infty$*, and the sequence is asymptotically powerless if*

$$\mathbb{P}_{\mathbb{H}_{0,n}}\{T_n \text{ rejects } \mathbb{H}_{0,n}\} + \mathbb{P}_{\mathbb{H}_{1,n}}\{T_n \text{ accepts } \mathbb{H}_{0,n}\} \to 1,$$

as $n \to \infty$.

For small $\eta > 0$, define the test that if $X_n^* > \sqrt{2(1+\eta)\log(n)}$, we reject $\mathbb{H}_{0,n}$; else accept $\mathbb{H}_{0,n}$. It is shown in [5] that for any $\eta > 0$ if $A_n = \sqrt{2(1+\eta)\log(n)}$, then the test is an asymptotically powerful statistic to detect the interval $[a, b)$ and there is algorithm using dyadic interval approximation running in $O(n)$ time with the ability to detect such an interval. But if $A_n \leq \sqrt{2(1-\eta)\log(n)}$, every sequence of tests (T_n) is asymptotically powerless. When $m = 1$, in our detection problem (5.1.44) versus (5.1.45), A_n is equal to $\mu \times \sqrt{|\mathcal{L}_n^0|}$ and hence if $\mu \times \sqrt{|\mathcal{L}_n^0|} \leq \sqrt{2(1-\eta)\log(n)}$, every sequence of tests (T_n) is asymptotically powerless and therefore in this case the elevated area is not detectable. In general case $(m > 1)$, one can see that if $\mu \times \sqrt{|\mathcal{L}_n^0|} \leq \sqrt{2(1-\eta)\log(n)}$, then no test will be able to detect these objects unless additional information regarding the object is known.

In the next section, we will show that when $|\mathcal{L}_n^0| \leq C\log(n)$ for a particular $C > 0$, our statistic and the related algorithm can compete with that in [5].

5.3 Likelihood ratio based approach

The following is an approach that first appeared in [5] to detect elevated interval on a line and regular embedded shapes such as rectangles and disks in a noisy image.

In [40], this approach is used to detect convex sets. The essence of this approach is to use a set with much less cardinality to approximate the set of objects which we are interested in. For example in [5], dyadic intervals and beamlets are employed to approximate regular intervals on a line and line segment respectively in a noisy image. In [40], the author uses a technique named hv-parallelograms to approximate the set of all planar bounded convex sets which strongly reduces the number of sets that are under consideration. The analysis is based on an asymptotic viewpoint such as the number of pixels goes to infinity. We find that this approach can be successfully applied in our problem to detect chains with good continuation if under $\mathbb{H}_1(\mathcal{L}_n^0, \mu)$, the length of the chain $|\mathcal{L}_n^0|$ is less than some multiple of $\log n$. Let us first consider a simple case. If we assume that both the position $\mathcal{L}_n^0 \subset \mathcal{S}$ and the mean $\mu > 0$ of the possible chain under $\mathbb{H}_1$ of (5.1.45) are given, we have a simple null hypothesis versus a simple alternative. Define

$$X(\mathcal{L}_n^0) = \sum_{(i,j)\in\mathcal{L}_n^0} \frac{X(i,j)}{\sqrt{|\mathcal{L}_n^0|}}, \tag{5.3.47}$$

where $|\mathcal{L}_n^0|$ is the number of nodes (pixels) in the chain $\mathcal{L}_n^0$. Therefore, under $\mathbb{H}_0$, it is not hard to obtain that $X(\mathcal{L}_n^0) \sim N(0,1)$, while under $\mathbb{H}_1(\mathcal{L}_n^0, \mu)$, we have $X(\mathcal{L}_n^0) \sim N(\mu\sqrt{|\mathcal{L}_n^0|}, 1)$. Thus, one can easily implement the likelihood ratio test of $\mathbb{H}_0$ against $\mathbb{H}_1(\mathcal{L}_n^0, \mu)$ by testing whether $X(\mathcal{L}_n^0) > \tau$, for a threshold τ.

For the composite alternative hypothesis, where both μ and $\mathcal{L}_n^0$ are unknown, it is straightforward to take the maximum of (5.3.47) among all possible chains. That is, we consider

$$X^* = \max_{\mathcal{L}\in\mathcal{F}_n} X(\mathcal{L}), \tag{5.3.48}$$

where $\mathcal{F}_n$ denotes the collection of all the subsets in $\mathcal{S}$ that are under consideration. To be specific, in our detection problem, let $\mathcal{F}_n$ be the set consisting of all chains with good continuation in the array, where the specific formulation of a chain is given in (5.1.43). Now we derive a detection rule so that for the composite alternative

$\mathbb{H}_1(\mathcal{L}_n^0, \mu)$, the probability of the type-I error (i.e., $\mathbb{P}(\text{accept}\mathbb{H}_1(\mathcal{L}_n^0, \mu)\big|\mathbb{H}_0)$) converges to 0 as the number of columns n goes to infinity. Consider a constant $\tau > 0$, under $\mathbb{H}_0$ by using the property of standard normal distribution (see p. 191 of [57]), we have

$$\forall \mathcal{L} \in \mathcal{F}_n, \mathbb{P}(X(\mathcal{L}) > \tau) < \frac{1}{\tau} \exp\{-\frac{\tau^2}{2}\}; \tag{5.3.49}$$

and

$$\mathbb{P}(X^* > \tau) \leq |\mathcal{F}_n| \cdot \mathbb{P}(X(\mathcal{L}) > \tau) \leq \frac{|\mathcal{F}_n|}{\tau} \exp\{-\frac{\tau^2}{2}\}, \tag{5.3.50}$$

where the first inequality in (5.3.50) is due to Bonferroni and $|\mathcal{F}_n|$ is the cardinality of the collection $|\mathcal{F}_n|$.

Note that in (5.3.50), if $\tau^* = \sqrt{2 \log |\mathcal{F}_n|} \to \infty$, then under $\mathbb{H}_0$, one has

$$\mathbb{P}(X^* > \tau^*) \to 0.$$

This gives us a powerful hypothesis testing method since the probability of the type-I error of this test goes to zero. On the other hand, considering a chain $\mathcal{L}$ within which there is a positive mean μ, we have $X(\mathcal{L}) \sim N(\mu\sqrt{|\mathcal{L}|}, 1)$. If the mean of this normal distribution $\mu\sqrt{|\mathcal{L}|} > \tau^*$ (respectively, $\mu\sqrt{|\mathcal{L}|} < \tau^*$), such a chain will (respectively, will *not*) be distinguished from the null hypothesis. Hence the aforementioned choice of the threshold $\tau^* = \sqrt{2 \log |\mathcal{F}_n|}$ gives an asymptotically powerful threshold on when a chain is detectable as $n \to \infty$.

However, for such a testing approach to be useful in determining the asymptotic detectability of chains with good continuation, the size of the collection $|\mathcal{F}_n|$ should grow slower than some polynomial expression of n. If $|\mathcal{F}_n| = O(n^k)$ (or $\lim_{n\to\infty} \mathcal{F}_n/n^k = C_1$), then $\tau^* \approx C_2\sqrt{2k \log n}$, where C_1 and C_2 are positive constants. Note that to increase the value of τ^* by a factor of 10, the value of n needs to be increased to n^{100}. The slow growth of τ^* when $|\mathcal{F}_n|$ is a polynomial is an interesting feature of this type of detection problems. In summary, the existence of a polynomial formula for the quantity $|\mathcal{F}_n|$ is of strong interest to us.

5.4 The infeasibility and the remedy

We have clarified the importance of the cardinality of the collection of all possible chains in the array. However the aforementioned proposal is infeasible due to the fast growth of the collection.

Theorem 5.4.1. *Under the definition of an embedded chain, unless $m = 1$ or $C = 0$, the number of all possible chains increases faster than any finite degree polynomial of n, as $n \to \infty$.*

Proof of Theorem 5.4.1. It suffices to show the proof when $m = 2$ and $C = 1$ since the bigger m and C will give a large array and hence more chains. Consider all the possible chains starting at column 1 and ending up at column n. It is easy to see that we will have 2^n such chains. Therefore $|\mathcal{F}_n| > 2^n$ has an exponential growth of n. □

This result implies that the approach we introduced in the previous section for determining the asymptotic threshold of the detectability of embedded chains cannot work. However, we would like to point out that, even though $|\mathcal{F}_n|$ is not a polynomial of n, if $|\mathcal{L}_n^0|$ is small comparing to n under $\mathbb{H}_1(\mathcal{L}_n^0, \mu)$ it would still be possible to derive a threshold $\tau^* \approx C_1\sqrt{2\log n}$ which can separate $\mathbb{H}_0$ from $\mathbb{H}_1(\mathcal{L}_n^0, \mu)$.

For a constant $0 < \zeta < \frac{1}{\log(2C+1)} \leq 1$, let $\mathcal{F}_n^t$ be a subset of $\mathcal{F}_n$ such that all the chains in $\mathcal{F}_n^t$ is no longer than $\zeta \log n$, i.e.,

$$\mathcal{F}_n^t = \{\mathcal{L} \in \mathcal{F}_n : |\mathcal{L}| \leq \zeta \log n\}. \tag{5.4.51}$$

The value of $\frac{1}{\log(2C+1)}$ is shown in the table below.

Table 2: The value of $1/\log(2C+1)$.

C	1	2	3	4	5
$1/\log(2C+1)$	0.9102	0.6213	0.5139	0.4551	0.4170

One can derive the following upper bound of $|\mathcal{F}_n^t|$,

$$\sum_{k=0}^{\zeta \log n} mn(2C+1)^k \leq mn(2C+1) \cdot (2C+1)^{\zeta \log n} \leq (2C+1)mn^2,$$

where the last inequality holds since $\zeta < \frac{1}{\log(2C+1)}$. It follows that

$$\sqrt{2\log|\mathcal{F}_n^t|} \leq \sqrt{2\log[mn^2(2C+1)]} = \sqrt{2(\log m + 2\log n + \log(2C+1))}.$$

Note that $\log m \ll \log n$ and $\log(2C+1) \ll \log n$ as n becomes large and it follows that there is a small $\eta > 0$ such that

$$\sqrt{2\log|\mathcal{F}_n^t|} \leq \sqrt{4(1+\eta)\log n} \quad \text{as} \quad n \to \infty.$$

Let $\tau_t^* = \sqrt{4(1+\eta)\log n}$ and let X_t^* be the statistic in (5.3.48) over $\mathcal{F}_n^t$, i.e.,

$$X_t^* = \max_{\mathcal{L} \in \mathcal{F}_n^t} \sum_{(i,j)\in\mathcal{L}} \frac{X(i,j)}{\sqrt{|\mathcal{L}|}}. \tag{5.4.52}$$

Note by (5.3.49) and (5.3.50), under the null hypothesis $\mathbb{H}_0$, one can immediately see that

$$\mathbb{P}(X_t^* > \tau_t^*) \leq \left|\mathcal{F}_n^t\right| \exp\{-\frac{(\tau_t^*)^2}{2}\}/\tau_t^* \to 0, \quad \text{as} \quad n \to \infty. \tag{5.4.53}$$

It takes $O(n\log n)$ flops to calculate the statistic X_t^* and we summarize our result in the next theorem.

Theorem 5.4.2. *In an array $\mathcal{S}$ of m-by-n nodes, consider the following detection problem*

$$\mathbb{H}_0 : X(i,j) \sim N(0,1), i.i.d., \forall (i,j) \in \mathcal{S}$$

versus

$$\mathbb{H}_1(\mathcal{L}_n^0, \mu) : X(i,j) \sim \mu + N(0,1), i.i.d., \forall (i,j) \in \mathcal{L}_n^0,$$

where $\mathcal{L}_n^0$ is a chain with good continuation (C apart). For constant $0 < \zeta < \frac{1}{\log(2C+1)}$, assume that under $\mathbb{H}_1(\mathcal{L}_n^0, \mu)$ the length of $\mathcal{L}_n^0$ is no longer than $\zeta\log(n)$ as $n \to \infty$. Define a statistic

$$X_t^* = \max_{\mathcal{L} \in \mathcal{F}_n^t} \sum_{(i,j)\in\mathcal{L}} \frac{X(i,j)}{\sqrt{|\mathcal{L}|}},$$

where $\mathcal{F}_n^t$ defined in (5.4.51), is the subset of $\mathcal{F}_n$. Then for any small $\eta > 0$, let the threshold τ_t^* be $\sqrt{4(1+\eta)\log n}$, then we have

$$\mathbb{P}(X_t^* > \tau_t^* \big| \mathbb{H}_0) \to 0, \quad as \quad n \to \infty. \tag{5.4.54}$$

If $\mu\sqrt{|\mathcal{L}_n^0|} > \tau_t^*$, then we have

$$\mathbb{P}(X_t^* < \tau_t^* \big| \mathbb{H}_1(\mathcal{L}_n^0, \mu)) \to 0, \quad as \quad n \to \infty. \tag{5.4.55}$$

Hence under the above condition, the test based on X_t^* is asymptotically powerful. Finally, given a realization of variables $\{X(i,j) : 1 \le m, 1 \le j \le n\}$, there is a dynamic programming algorithm for computing the value of X_t^* and the computational time is upper-bounded by $C_1 n \log n$ for some $C_1 > 0$ and the required space is also $O(n \log n)$.

Proof of Theorem 5.4.2. We showed (5.4.54) in (5.4.53) and it suffices to prove (5.4.55) and the complexity of the algorithm. Since $|\mathcal{L}_n^0| \le \zeta \log n$ by our assumption, $\mathcal{L}_n^0 \in \mathcal{F}_n^t$ and therefore under $\mathbb{H}_1(\mathcal{L}_n^0, \mu)$ we have

$$X_t^* \ge \sum_{(i,j)\in\mathcal{L}_n^0} \frac{X(i,j)}{\sqrt{|\mathcal{L}_n^0|}} \sim N(\mu\sqrt{|\mathcal{L}_n^0|}, 1).$$

Hence under the alternative hypothesis by Mills' Ratio

$$\begin{aligned}
\mathbb{P}\left[X_t^* < \tau_t^* \big| \mathbb{H}_1(\mathcal{L}_n^0, \mu)\right] &\le \mathbb{P}\left[N(\mu\sqrt{|\mathcal{L}_n^0|}, 1) < \tau_t^* \big| \mathbb{H}_1(\mathcal{L}_n^0, \mu)\right] \\
&\le \mathbb{P}(N(0,1) < -\gamma \log n) \\
&\le 2n^{-\gamma^2/2} \to 0, \quad \text{as} \quad n \to \infty,
\end{aligned}$$

where $\gamma = \mu\sqrt{|\mathcal{L}_n^0|}/\sqrt{\log n} - \sqrt{4(1+\eta)} > 0$ by our assumption. Let us see the algorithm which takes $O(n \log n)$ flops to compute the value of X_t^*. Denote $\zeta \log n$ by S. Given a realization $\{X(i,j) : 1 \le i \le m, 1 \le j \le n\}$, let Y be

$$\{Y(i,j,s) : 1 \le i \le m, 1 \le j \le n, 1 \le s \le S\}$$

such that

$$\begin{aligned} Y(i,j,1) &= X(i,j) \\ Y(i,j,s) &= X(i,j) + \max_{i' \in \Omega(i)} Y(i', j-1, s-1), \quad \text{for} \quad s = 2, \ldots, \min\{j, S\}; \\ Y(i,j,s) &= -3mn, \quad \text{otherwise}, \end{aligned}$$

where $\Omega(i) = \{i' : |i' - i| \leq C, 1 \leq i' \leq m\}$ denotes the set containing neighboring indices of i. It is easy to see that

$$X_t^* = \max_{1 \leq i \leq m, 1 \leq j \leq n, 1 \leq s \leq S} Y(i,j,s)/\sqrt{s},$$

and it takes $O(n \log n)$ to compute Y and hence X_t^*. □

Under $\mathbb{H}_1(\mathcal{L}_n^0, \mu)$ with $|\mathcal{L}_n^0| > \zeta \log n$, it is not hard to realize that (5.4.55) may fail because $\mathcal{L}_n^0$ might be longer than $\zeta \log n$ and thus does not belong to $\mathcal{F}_n^t$. We introduce a longest significant run approach below to handle this case.

5.5 A longest significant run approach

In this section, we introduce a method which can separate $\mathbb{H}_0$ from $\mathbb{H}_1(\mathcal{L}_n^0, \mu)$ with $|\mathcal{L}_n^0| \geq \zeta \log n$. This test was independently considered in a series of papers ([45, 48]). Throughout this section, we assume that $|\mathcal{L}_n^0| \to \infty$ as $n \to \infty$ under $\mathbb{H}_1(\mathcal{L}_n^0, \mu)$. Given a prescribed threshold x^*, let us say a node $(i,j) \in \mathcal{S}$ is significant if its corresponding normal random variable $X(i,j) > x^*$. Recall that $\mathcal{S} = \{(i,j) : 1 \leq i \leq n, 1 \leq j \leq m\}$ is the set of pixels in the image and we define a random variable $z : \mathcal{S} \to \{0,1\}$ to indicate the significance of nodes, i.e., $z(i,j) = 1$ if $X(i,j) > x^*$ and $z(i,j) = 0$ otherwise. It is easy to see that under the null hypothesis, $z(i,j)$ has i.i.d Bernoulli distribution with success probability

$$p = \mathbb{P}(N(0,1) > x^*).$$

Given this notation, a chain with good continuation $\mathcal{L} \in \mathcal{F}_n$ is said to be significant if all the nodes on the chain are significant. Define a random variable $Z : \mathcal{F}_n \to \{0,1\}$

such that $Z(\mathcal{L}) = 1$ if $\mathcal{L} \in \mathcal{F}_n$ is significant and $Z(\mathcal{L}) = 0$ otherwise. Obviously the probability of a chain $\mathcal{L}$ to be significant under $\mathbb{H}_0$ is the following,

$$\mathbb{P}(Z(\mathcal{L}) = 1) = p^{|\mathcal{L}|}. \tag{5.5.56}$$

Given the aforementioned notation, let us recall a series of results in [11] which state the asymptotic rate of the length of the longest significant chain with good continuation in the m-by-n array of nodes as $n \to \infty$. We restate these results in Theorems 3.2.1, 3.2.3, 3.2.5 and Lemma 3.2.2 and Corollary 3.2.4 in Section 3.2.1 . Table 3 gives the exact values of ρ defined in Lemma 3.2.2 for different p's and m's: $m = 4, 8, 10$ and $C = 1$. The algorithmic complexity for finding ρ is $O(2^{3m})$.

Table 3: The values of ρ for different values of m and p, when $C = 1$.

p	0.1	0.2	0.3	0.4	0.5	0.6
m=4	0.2444	0.4564	0.6341	0.7758	0.8804	0.9482
m=8	0.2654	0.4955	0.6869	0.8363	0.9383	0.9876
m=10	0.2691	0.5022	0.6958	0.8467	0.9486	0.9930

Let $L_0(n)$ be the longest significant chain in the image and let $|L_0(n)|$ be its length. Notice that $|L_0(n)|$ actually depends on m, C and p in addition to n but we simplify the notation because all the parameters except for n are constant. By Theorem 3.2.3 and Egoroff's Theorem (See [62]), given any small $\epsilon > 0$ and $\delta > 0$, there is a large $N \in \mathbb{Z}^+$ such that for all $n \geq N$ with probability $1 - \delta$ under $\mathbb{H}_0$ we have

$$\left| \frac{|L_0(n)|}{\log_{1/\rho} n} - 1 \right| < \epsilon. \tag{5.5.57}$$

Denote $(1 + \epsilon) \log_{1/\rho} n$ by $b_{\epsilon,n}$, which under $\mathbb{H}_0$ is the upper bound of the length of the longest significant run with probability $1 - \delta$. Let $\mathcal{E}_n$ be a random subset of $\mathcal{F}_n$ which consists of all significant chains, i.e.,

$$\mathcal{E}_n = \{\mathcal{L} \in \mathcal{F}_n : Z(\mathcal{L}) = 1\}. \tag{5.5.58}$$

It is not hard to see that under the null hypothesis $\mathbb{H}_0$ for large $n \geq N$, with probability $1-\delta$, an upper bound of $|\mathcal{E}_n|$ under H_0 is

$$\sum_{k=1}^{b_{\epsilon,n}+1} mn(2C+1)^{k-1} = mn[(2C+1)^{b_{\epsilon,n}+1} - 1]/2C \tag{5.5.59}$$

By Corollary 3.2.4, we can choose the threshold of significant x^* large, so that

$$p = \mathbb{P}(N(0,1) > x^*)$$

is small enough to render $\rho < \frac{1}{2C+1}$. By Corollary 3.2.4, it is easy to see that $p \leq \rho < \frac{1}{2C+1}$. Thus an upper bound on $|\mathcal{E}_n|$ is

$$mn\frac{2C+1}{2C} \cdot (2C+1)^{\log_{2C+1} n\cdot(1+\epsilon)\log_{1/\rho}(2C+1)} \leq mn\frac{2C+1}{2C} n^{(1+\epsilon)\log_{1/\rho}(2C+1)} \tag{5.5.60}$$

Since ϵ is arbitrary, as n becomes large, the right hand side of (5.5.60) is less than $mn^{2-\delta_0}$ for some $\delta_0 > 0$. Thus under the null hypothesis $\mathbb{H}_0$, $|\mathcal{E}_n|$ grows slower than n^2.

Now, we define our statistic based on all significant chains in the array $\mathcal{S}$.

Definition 5.5.1. *In an array of m-by-n nodes $\mathcal{S}$, let $X(i,j)$ be the normal random variable associated with each node (i,j). Let $X(\mathcal{L})$ be*

$$\sum_{(i,j)\in\mathcal{L}} \frac{X(i,j)}{\sqrt{|\mathcal{L}|}}.$$

Then define a statistic based on all significant chains to be

$$X_s^* = \max_{\mathcal{L}\in\mathcal{E}_n} X(\mathcal{L}). \tag{5.5.61}$$

Given the aforementioned notation of x^*, $\mathcal{E}_n$ and ρ, we now describe the algorithm for the analysis of a noisy image $\mathcal{S} = \{X(i,j), 1 \leq i \leq m, 1 \leq j \leq n\}$ looking for suspected chains with good continuation in $\mathcal{S}$. The algorithm has several steps and its computational complexity is $O(n \log n)$.

- **Step I:** Count the length of the longest chain $L_0(n)$ in $\mathcal{E}_n$. If the length

$$|L_0(n)| > (1+\epsilon/2)\log_{1/\rho} n$$

for some small $\epsilon > 0$, then reject $\mathbb{H}_0$; otherwise, go to Step II.

- **Step II:** Compute X_s^* as in (5.5.61). If $X_s^* > \sqrt{2(1+\delta_2)\log n}$ for some small $\delta_2 > 0$, then reject $\mathbb{H}_0$; otherwise, accept $\mathbb{H}_0$.

We show the algorithms to find $|L_0(n)|$ and X_s^* below.

Algorithm to find $|L_0(n)|$: Recall that for a node $(i,j) \in \mathcal{S}$, we use $z(i,j) = 1(=0)$ to denote the (in)significance of (i,j). Given a realization $\{X(i,j) : 1 \le i \le m, 1 \le j \le n\}$, let Y_1 be $\{Y_1(i,j) : 1 \le i \le m, 1 \le j \le n\}$ such that

$$\begin{aligned} Y_1(i,1) &= z(i,1), i = 1,\ldots,m \\ Y_1(i,j) &= z(i,j)(1 + \max_{i' \in \Omega(i)} Y_1(i', j-1)), 1 = 1,\ldots,m, j = 2,\ldots,n, \end{aligned}$$

where $\Omega(i) = \{i' : |i'-i| \le C, 1 \le i' \le m\}$ denotes the set containing neighboring indices of i. Finally, let $|L_0(n)|$ be

$$\max_{(i,j)\in\mathcal{S}} Y(i,j).$$

It is not hard to see that this algorithm takes Cmn time for $C > 0$.

Algorithm to find X_s^*: Let $S = 3\log_{1/\rho} n$, $S = 3\log_{1/\rho} n$ and Y_2 be $\{Y_2(i,j,s) : 1 \le i \le m, 1 \le j \le n, 1 \le s \le S\}$ such that

$$\begin{aligned} Y_2(i,j,1) &= X(i,j); i = 1,\ldots,m, j = 1\ldots,n; \\ Y_2(i,j,s) &= z(i,j)(X(i,j) + \max_{i' \in \Omega(i)} Y_2(i', j-1, s-1)), \\ &\quad i = 1\ldots,m, j = 2\ldots,n, 2 \le s \le 3\log_{1/\rho} n; \\ X_s^* &= \max_{(i,j)\in\mathcal{S}} \frac{Y_2(i,j,s)}{\sqrt{s}}, \end{aligned}$$

where $\Omega(i)$ is the set of neighboring indices of i as above. It is not hard to see that this algorithm takes $Cmn \log n$ time for some $C > 0$.

The next two subsections give an upper bound of the probability of type-I error under $\mathbb{H}_0$ and type-II error under $\mathbb{H}_1(\mathcal{L}_n^0, \mu)$.

5.5.1 Behavior under $\mathbb{H}_0$.

We need to show that with overwhelming probability under $\mathbb{H}_0$, there will be no significant chain longer than $(1+\epsilon)\log_{1/\rho} n$ or $X_s^* < \sqrt{2(1+\delta_2)\log n}$ for small $\epsilon > 0$ and $\delta_2 > 0$ as $n \to \infty$. The former is shown in (5.5.57) and we show the latter in the next theorem.

Theorem 5.5.2. *Under the null hypothesis $\mathbb{H}_0$, for any small $\delta > 0$, there exists a constant σ_1 depending on p and a large $N \in \mathbb{Z}^+$ such that for any $n \geq N$ we have the following:*

$$\mathbb{P}(X_s^* > \tau_s^*) \leq mn\sigma_1 \exp\{-\frac{(\tau_s^*)^2}{2}\} + \delta. \tag{5.5.62}$$

Thus for any $\delta_2 > 0$, when $\tau_s^* = \sqrt{2(1+\delta_2)\log n}$, we have $\mathbb{P}_{\mathbb{H}_0}(X_s^* > \tau_s^*) \to 0$ as $n \to \infty$.

Proof of Theorem 5.5.2. Recall (5.5.57) that for any $\delta > 0$ and $\epsilon > 0$, the length of the longest significant chain under $\mathbb{H}_0$ is no larger than $b_{\epsilon,n} = (1+\epsilon)\log_{1/\rho} n$ with probability $1-\delta$. By the Bonferroni inequality, it is easy to derive the following,

$$\begin{aligned}
& \mathbb{P}(X_s^* > \tau_s^*) \\
= \; & \mathbb{P}(\bigcup_{\mathcal{L}\in\mathcal{F}_n} \{X(\mathcal{L}) > \tau_s^*, Z(\mathcal{L}) = 1\}) \\
\leq \; & \delta + \mathbb{P}(\bigcup_{k=1}^{b_{\epsilon,n}} \bigcup_{\mathcal{L}\in\mathcal{F}_n, |\mathcal{L}|=k} \{X(\mathcal{L}) > \tau_s^*, Z(\mathcal{L}) = 1\}) \\
\leq \; & \delta + \sum_{k=1}^{b_{\epsilon,n}} \sum_{\mathcal{L}\in\mathcal{F}_n, |\mathcal{L}|=k} \mathbb{P}(X(\mathcal{L}) > \tau_s^* \big| Z(\mathcal{L}) = 1) \cdot \mathbb{P}(Z(\mathcal{L}) = 1) \\
\leq \; & \delta + \sum_{k=1}^{b_{\epsilon,n}} \sum_{\mathcal{L}\in\mathcal{F}_n, |\mathcal{L}|=k} p^k \mathbb{P}(X(\mathcal{L}) > \tau_s^* \big| Z(\mathcal{L}) = 1)
\end{aligned} \tag{5.5.63}$$

where $|\mathcal{L}|$ is the length of the significant chain $\mathcal{L}$. Note that conditioning on the event $Z(\mathcal{L}) = 1$, each $X(i,j)$ on $\mathcal{L}$ is a truncated standard normal random variable bounded below by x^*. Let $X_T(i,j)$ $(1 \le i \le n, 1 \le j \le m)$ be i.i.d. random variables with distribution equal to Z given that $Z > x^*$, where Z is the standard normal random variable. Let $\Phi(\cdot)$ be the distribution function of the standard normal distribution. Thus for each $\mathcal{L} \in \mathcal{E}_n$, we have

$$
\begin{aligned}
& \mathbb{P}(X(\mathcal{L}) > \tau_s^* \big| Z(\mathcal{L}) = 1) \\
= \; & \mathbb{P}(\sum_{(i,j)\in\mathcal{L}} \frac{X(i,j)}{\sqrt{|\mathcal{L}|}} > \tau_s^* \big| Z(\mathcal{L}) = 1) = \mathbb{P}(\sum_{(i,j)\in\mathcal{L}} X_T(i,j) > \tau_s^* \sqrt{|\mathcal{L}|}) \\
\le \; & \inf_{\omega\ge 0} \exp\{-\omega\tau_s^*\sqrt{|\mathcal{L}|}\} \prod_{(i,j)\in\mathcal{L}} \mathbb{E}\exp\{\omega X_T(i,j)\} \\
\le \; & \inf_{\omega\ge 0} \exp\{-\omega\tau_s^*\sqrt{|\mathcal{L}|}\} \prod_{(i,j)\in\mathcal{L}} \frac{1}{(1-\Phi(x^*))\sqrt{2\pi}} \int_{x^*}^{\infty} \exp\{\omega y - \frac{y^2}{2}\}\mathrm{d}y \\
= \; & \inf_{\omega\ge 0} \exp\{-\omega\tau_s^*\sqrt{|\mathcal{L}|}\} \prod_{(i,j)\in\mathcal{L}} \frac{\exp\{\frac{\omega^2}{2}\}}{(1-\Phi(x^*))\sqrt{2\pi}} \int_{x^*}^{\infty} \exp\{-\frac{(y-\omega)^2}{2}\}\mathrm{d}y \\
= \; & \inf_{\omega\ge 0} \exp\{-\omega\tau_s^*\sqrt{|\mathcal{L}|}\} \prod_{(i,j)\in\mathcal{L}} \frac{\exp\{\frac{\omega^2}{2}\}}{(1-\Phi(x^*))\sqrt{2\pi}} \int_{x^*+\omega}^{\infty} \exp\{-\frac{z^2}{2}\}\mathrm{d}z \\
= \; & \inf_{\omega\ge 0} \exp\{-\omega\tau_s^*\sqrt{|\mathcal{L}|}\} \prod_{(i,j)\in\mathcal{L}} \frac{\exp\{\frac{\omega^2}{2}\}}{(1-\Phi(x^*))} (1-\Phi(x^*+\omega)) \\
\le \; & \inf_{\omega\ge 0} \exp\{-\omega\tau_s^*\sqrt{|\mathcal{L}|} + |\mathcal{L}|\,\frac{\omega^2}{2}\} \\
= \; & \inf_{\omega\ge 0} \exp\{\frac{1}{2}(\omega\sqrt{|\mathcal{L}|} - \tau_s^*)^2 - \frac{(\tau_s^*)^2}{2}\} \\
\le \; & \exp\{-\frac{(\tau_s^*)^2}{2}\}. \qquad (5.5.64)
\end{aligned}
$$

Plug (5.5.64) into (5.5.63), since $p = \mathbb{P}(N(0,1) > x^*) \le \rho < \frac{1}{2C+1}$ we have

$$
\begin{aligned}
\mathbb{P}(X_s^* > \tau_s^*) \; & \le \; \delta + \sum_{k=1}^{b_{\epsilon,n}} \sum_{\mathcal{L}\in\mathcal{F}_n, |\mathcal{L}|=k} p^k \exp\{-\frac{(\tau_s^*)^2}{2}\} \\
& \le \; \delta + mn\exp\{-\frac{(\tau_s^*)^2}{2}\} \sum_{k=1}^{b_{\epsilon,n}} (2C+1)^k p^k \\
& \le \; \delta + mn\frac{1}{1-(2C+1)p}\exp\{-\frac{(\tau_s^*)^2}{2}\}.
\end{aligned}
$$

Since $\delta > 0$ is arbitrary and m is fixed, if $\tau_s^* = \sqrt{(2+\delta_2)\log mn}$, then we have that $\mathbb{P}(X_s^* > \tau_s^* \big| \mathbb{H}_0) \to 0$ for any small $\delta_2 > 0$, as $n \to \infty$. Since m is fixed, $\log m \ll \log n$ as $n \to \infty$, $\tau_s^* = \sqrt{2(1+\delta_2)\log n}$ eventually. □

So far, we have studied the asymptotic behavior of X_s^* under $\mathbb{H}_0$ and proved that the type-I error tends to 0 as $n \to \infty$. In the next subsection, we delve into the behavior of X_s^* under $\mathbb{H}_1(\mathcal{L}_n^0, \mu)$ and shall prove the diminishing type-II error.

5.5.2 Asymptotic behavior under $\mathbb{H}_1(\mathcal{L}_n^0, \mu)$

We first show the condition under which the type-II error diminishes fast under $\mathbb{H}_1(\mathcal{L}_n^0, \mu)$ if $|\mathcal{L}_n^0| \geq \zeta_1 n$. Let us first see the behavior of the longest significant chain embedded in $\mathcal{L}_n^0$ under $\mathbb{H}_1(\mathcal{L}_n^0, \mu)$. Denote $\mathbb{P}(N(\mu, 1) > x^*)$ by p_1 which is the probability of nodes to be significant in the chain $\mathcal{L}_n^0$. Let $L_1(n)$ be the longest significant chain in $\mathcal{L}_n^0$ and $|L_1(n)|$ be its length. Recall that $|L_0(n)|$ is the length of the longest significant chain in the image and so $|L_1(n)| \leq |L_0(n)|$ since $L_1(n) \in \mathcal{E}_n$. As $n \to \infty$ and hence $|\mathcal{L}_n^0| \to \infty$, by Theorem 3.2.3, we have the following convergence rate of $|L_1(n)|$,

$$\frac{|L_1(n)|}{\log_{1/p_1} |\mathcal{L}_n^0|} \to 1 \quad \text{almost surely.} \tag{5.5.65}$$

Therefore by Ergoroff's theorem, given any small $\epsilon_1 > 0$ and $\delta > 0$, there exists a large $N \in \mathbb{Z}^+$ such that for all $n \geq N$ with probability $1 - \delta$ we have

$$\left| \frac{|L_1(n)|}{\log_{1/p_1} |\mathcal{L}_n^0|} - 1 \right| < \epsilon. \tag{5.5.66}$$

That is to say $(1-\epsilon_1)\log_{1/p_1} |\mathcal{L}_n^0| \leq L_1(n) \leq (1+\epsilon_1)\log_{1/p_1} |\mathcal{L}_n^0|$ with probability at least $1 - \delta$. We will prove the following theorem regarding the type-II error.

Theorem 5.5.3. *In an array $\mathcal{S}$ of m-by-n nodes, consider the following detection problem*

$$\mathbb{H}_0 : X(i,j) \sim N(0,1), i.i.d., \forall (i,j) \in \mathcal{S}$$

versus

$$\mathbb{H}_1(\mathcal{L}_n^0, \mu) : X(i,j) \sim \mu + N(0,1), i.i.d., \forall (i,j) \in \mathcal{L}_n^0,$$

where $\mathcal{L}_n^0$ is a chain with good continuation (C apart). Assume that the length of $\mathcal{L}_n^0$ is at least $\zeta_1 n$ for some $\zeta_1 > 0$. If μ is such that

$$p_1 > \rho^{\log \zeta_1 n/(1+\epsilon)\log n} \to \rho^{1/(1+\epsilon)} \quad as \quad n \to \infty \tag{5.5.67}$$

for some small $\epsilon > 0$, then we have

$$\mathbb{P}(|L_0(n)| > (1+\epsilon/2)\log_{1/\rho} n \big| \mathbb{H}_1(\mathcal{L}_n^0, \mu)) \to 1, \quad as \quad n \to \infty. \tag{5.5.68}$$

Proof of Theorem 5.5.67. By the argument before the theorem, for any small $\delta > 0$ and $\epsilon_1 > 0$, with probability $1-\delta$, there exists $N \in \mathbb{Z}^+$ such that when $n \geq N$,

$$|L_0(n)| \geq |L_1(n)| \geq (1-\epsilon_1)\log_{1/p_1} \left|\mathcal{L}_n^0\right|.$$

Under $\mathbb{H}_1(\mathcal{L}_n^0, \mu)$, by (5.5.67), one can derive

$$\log_{1/p_1} \left|\mathcal{L}_n^0\right| \geq \log_{1/p_1} \zeta_1 n > (1+\epsilon)\log_{1/\rho} n.$$

We choose ϵ_1 such that $(1+\epsilon)(1-\epsilon_1) > 1+\epsilon/2$ and thus

$$\mathbb{P}[|L_0(n)| > (1+\epsilon/2)\log_{1/\rho} n \big| \mathbb{H}_1(\mathcal{L}_n^0, \mu)] \geq 1-\delta, \forall n \geq N.$$

Since δ is arbitrary, we have (5.5.68) asymptotically. The second part of (5.5.67) follows from the fact that

$$\lim_{n\to\infty} \frac{\log \zeta n}{(1+\epsilon)\log n} = \lim_{n\to\infty} \frac{\log \zeta + \log n}{(1+\epsilon)\log n} = \frac{1}{1+\epsilon}.$$

□

Let us consider the case of $\mathbb{H}_1(\mathcal{L}_n^0, \mu)$ such that $|L_0(n)| \leq (1+\epsilon/2)\log n$. Denote two constants $\lceil (1-\epsilon)\log_{1/p_1} |\mathcal{L}_n^0| \rceil$ and $\lfloor (1+\epsilon)\log_{1/p_1} |\mathcal{L}_n^0|$ by $c_{\epsilon,n}^L$ and $c_{\epsilon,n}^U$ respectively. Recall a chain $\mathcal{L}$ is in $\mathcal{E}_n$ if and only if $X(i,j) > x^*$ for every node $(i,j) \in \mathcal{L}$, where

x^* is the threshold of significance. In Definition 5.5.1, X_s^* is said to be the maximum of all $X(\mathcal{L})$ among $\mathcal{L} \in \mathcal{E}_n$, which is of course no smaller than $X(L_1(n))$. We now give the asymptotic diminishing rate of the type-II error.

Theorem 5.5.4. *Under the alternative hypothesis $\mathbb{H}_1(\mathcal{L}_n^0, \mu)$, for any small $\delta > 0$ and $\epsilon > 0$, if $\mu\sqrt{c^L(\epsilon, n)} > \tau_s^*$, we have the following:*

$$\mathbb{P}(X_s^* > \tau_s^* \big| \mathbb{H}_1(\mathcal{L}_n^0, \mu)) \to 1, \quad as \quad n \to \infty.$$

Before the proof, let us first see recall the following definition about the association of random variables in [28].

Definition 5.5.5. *We say random variables $T_1, T_2, \ldots, T_n$ are associated if*

$$Cov[f(\mathbf{T}), \mathbf{g}(\mathbf{T})] \geq \mathbf{0},$$

where $\mathbf{T} = (\mathbf{T_1}, \mathbf{T_2}, \ldots, \mathbf{T_n})$, for any nondecreasing functions f and g for which $\mathbb{E}f(\mathbf{T})$, $\mathbb{E}g(\mathbf{T})$, $\mathbb{E}f(\mathbf{T})\mathbf{g}(\mathbf{T})$ exist.

Now we give the proof of Theorem 5.5.4 in the following.

Proof of Theorem 5.5.4. Recalling (5.5.66), it is not difficult to derive the following,

$$\begin{aligned}
\mathbb{P}(X_s^* > \tau_s^*) &= \mathbb{P}(\bigcup_{\mathcal{L} \in \mathcal{F}_n} \{X(\mathcal{L}) > \tau_s^*, Z(\mathcal{L}) = 1\}) \\
&\geq \mathbb{P}(X(L_1(n)) > \tau_s^*) \\
&= \sum_{k=1}^{n} \mathbb{P}(\sum_{(i,j) \in L_1(n)} \frac{X(i,j)}{\sqrt{|L_1(n)|}} > \tau_s^* \big| \, |L_1(n)| = k) \mathbb{P}(|L_1(n)| = k) \\
&\geq \sum_{k=c_{\epsilon,n}^L}^{c_{\epsilon,n}^U} \mathbb{P}(\sum_{(i,j) \in L_1(n)} \frac{X(i,j)}{\sqrt{k}} > \tau_s^* \big| \, |L_1(n)| = k) \\
&\times \mathbb{P}(|L_1(n)| = k) \qquad (5.5.69)
\end{aligned}$$

Recall that $\mathcal{F}_n$ is the set of all chains of good continuation in

$$\mathcal{S} = \{(i,j) : 1 \leq i \leq n, 1 \leq j \leq m\}.$$

Let $\mathcal{F}_n^k \subset \mathcal{F}_n$ be the set of all chains of good continuation with length k, namely,

$$\mathcal{F}_n^k = \{\mathcal{L} \in \mathcal{F}_n : |\mathcal{L}| = k\}.$$

One can easily see that

$$\begin{aligned} & \mathbb{P}\Big(\sum_{(i,j)\in L_1(n)} \frac{X(i,j)}{\sqrt{k}} > \tau_s^* \Big| \, |L_1(n)| = k\Big) \\ = & \sum_{\mathcal{L}\in\mathcal{F}_n^k} \mathbb{P}\Big(\sum_{(i,j)\in\mathcal{L}} \frac{X(i,j)}{\sqrt{k}} > \tau_s^* \Big| L_1(n) = \mathcal{L}\Big)\mathbb{P}(L_1(n) = \mathcal{L}) \Big/ \sum_{\mathcal{L}\in\mathcal{F}_n^k} \mathcal{P}(L_1(n) = \mathcal{L}). \end{aligned}$$

Generate k random variables $Y_1, \ldots, Y_k \overset{\text{i.i.d.}}{\sim} N(\mu, 1)$ which are independent from $\{X_{i,j}, 1 \le i \le n, 1 \le j \le m\}$. For each $\mathcal{L} \in \mathcal{F}_n^k$, we have that

$$\begin{aligned} & \mathbb{P}\Big(\sum_{(i,j)\in\mathcal{L}} \frac{X(i,j)}{\sqrt{k}} > \tau_s^* \Big| L_1(n) = \mathcal{L}\Big) \\ = & \mathbb{P}\Big(\sum_{i=1}^k Y_i/\sqrt{k} > \tau_s^* \Big| Y_1 > x^*, \ldots, Y_k > x^*\Big) \end{aligned}$$

Let A and B be two subsets of $\mathbb{R}^k$ such that

$$A = \{(y_1, \ldots, y_k) : \sum_{i=1}^k \frac{y_i}{\sqrt{k}} > \tau_s^*\} \text{ and } B = \{(y_1, \ldots, y_k) : y_1 > x^*, \ldots, y_k > x^*\}.$$

Let $f : \mathbb{R}^k \to \{0, 1\}$ and $g : \mathbb{R}^k \to \{0, 1\}$ be indicator functions of sets A and B, respectively, i.e.,

$$f(y_1, \ldots, y_k) = I_A(y_1, \ldots, y_k) \text{ and } g(y_1, \ldots, y_k) = I_B(y_1, \ldots, y_k).$$

Theorem 2.1 of [28] states that independent random variables are associated. Therefore $Y_1, \ldots, Y_k$ are associated since they are independent. Realize that both f and g are increasing functions and therefore, it is straightforward to see that

$$\begin{aligned} & \mathbb{P}\Big(\sum_{i=1}^k Y_i/\sqrt{k} > \tau_s^*, Y_1 > x^*, \ldots, Y_k > x^*\Big) = \mathbb{E}[f(Y_1, \ldots, Y_k)g(Y_1, \ldots, Y_k)] \\ \ge & \mathbb{E}f(Y_1, \ldots, Y_k)\mathbb{E}g(Y_1, \ldots, Y_k) \\ = & \mathbb{P}\Big(\sum_{i=1}^k Y_i/\sqrt{k} > x^*\Big)\mathbb{P}(Y_1 > x^*, \ldots, Y_k > x^*). \end{aligned}$$

Hence it follows that

$$\begin{aligned} & \mathbb{P}(\sum_{i=1}^{k} Y_i/\sqrt{k} > \tau_s^* \big| Y_1 > x^*, \ldots, Y_k > x^*) \\ \geq \; & \mathbb{P}(\sum_{i=1}^{k} Y_i/\sqrt{k} > \tau_s^*) \\ = \; & \mathbb{P}(N(\sqrt{k}\mu, 1) > \tau_s^*). \end{aligned}$$

Since $k \geq c^L(\epsilon, n)$, as long as $\mu\sqrt{c^L(\epsilon, n)} > \tau_s^* = \sqrt{2(1+\delta_2)\log n}$, by Mill's ratio we have

$$\mathbb{P}(N(\sqrt{k}\mu, 1) < \tau_s^*) \leq \mathbb{P}(N(0,1) < -\gamma\sqrt{\log n}) \leq 2n^{-\gamma^2/2} \to 0, \text{ as } n \to 0,$$

where $\gamma = \sqrt{k}\mu/\sqrt{\log n} - \sqrt{2(1+\delta_2)} \geq \sqrt{c^L(\epsilon, n)}\mu/\sqrt{\log n} - \sqrt{2(1+\delta_2)} > 0$. Therefore, going back to (5.5.69), it follows that

$$\begin{aligned} \mathbb{P}(X_s^* > \tau_s^*) \; & \geq \; \sum_{k=c_{\epsilon,n}^L}^{c_{\epsilon,n}^U} (1 - 2n^{-\gamma^2/2}) \mathbb{P}(|L_1(n)| = k) \\ & \geq \; (1 - 2n^{-\gamma^2/2})(1 - \delta). \end{aligned}$$

Since $\delta > 0$ is arbitrary, as $n \to \infty$, we have $\mathbb{P}(X_s^* > \tau_s^* \big| \mathbb{H}_1(\mathcal{L}_n^0, \mu)) \to 1$ as $n \to \infty$ $\square$

Since ϵ in the aforementioned theorem is arbitrary, $c^L(\epsilon, n) \approx \log_{1/p_1} |\mathcal{L}_n^0|$, we may change the condition $\mu\sqrt{c^L(\epsilon, n)} > \tau_s^*$ in Theorem 5.5.4 to $\mu\sqrt{\log_{1/p_1} |\mathcal{L}_n^0|} > \tau_s^*$.

5.6 Summary of algorithms to detect chains with good continuation

In this section, we summarize the algorithms of our detecting method in chains with good continuation. In an array $\mathcal{S}$ of m-by-n nodes, consider the following detection problem

$$\mathbb{H}_0 : X(i,j) \sim N(0,1), i.i.d., \forall (i,j) \in \mathcal{S}$$

versus

$$\mathbb{H}_1(\mathcal{L}_n^0, \mu) : X(i,j) \sim \mu + N(0,1), i.i.d., \forall (i,j) \in \mathcal{L}_n^0,$$

where $\mathcal{L}_n^0$ is a chain with good continuation (C apart) and $\mu > 0$.

Detection Algorithms:

- For a constant $\zeta < \frac{1}{\log(2C+1)}$, let $\mathcal{F}_n^t = \{\mathcal{L} \in \mathcal{F}_n : |\mathcal{L}| \leq \zeta \log n\}$. Take

$$X_t^* = \max_{\mathcal{L} \in \mathcal{F}_n^t} \sum_{(i,j) \in \mathcal{L}} \frac{X(i,j)}{\sqrt{|\mathcal{L}|}}.$$

  For any small $\eta > 0$, if $X_t^* > \tau_t^* = \sqrt{4(1+\eta)\log n}$, reject $\mathbb{H}_0$; otherwise go to the next step.

- Take x^* such that $\rho < \frac{1}{2C+1}$ to be the threshold of nodes to be significant. Let $\mathcal{E}_n = \{\mathcal{L} \in \mathcal{F}_n : Z(\mathcal{L}) = 1\}$. Find the longest chain $L_0(n)$ in $\mathcal{E}_n$. For small $\epsilon > 0$ if the length $|L_0(n)| > (1+\epsilon/2)\log_{1/\rho} n$, then reject $\mathbb{H}_0$; otherwise, go to the next step.

- Computing X_s^* as

$$\max_{\mathcal{L} \in \mathcal{E}_n} \sum_{(i,j) \in \mathcal{L}} \frac{X(i,j)}{\sqrt{|\mathcal{L}|}}.$$

  For small $\delta_2 > 0$, if $X_s^* > \sqrt{2(1+\delta_2)\log n}$, then reject $\mathbb{H}_0$; otherwise accept $\mathbb{H}_0$.

As shown above, the first step takes $O(n \log n)$ flops, the second $O(n)$ and third $O(n \log n)$. Hence this algorithm takes $O(n \log n)$ flops in total with $O(n \log n)$ required space for storage. Moreover, under $\mathbb{H}_1(\mathcal{L}_n^0, \mu)$ if $|\mathcal{L}_n^0| < \zeta \log n$ and $\mu\sqrt{|\mathcal{L}_n^0|} > \sqrt{4(1+\eta)\log n}$; or if $|\mathcal{L}_n^0| \geq \zeta_1 n$ for some $\zeta_1 > 0$ as $n \to \infty$ and $\mathbb{P}(N(\mu,1) > x^*) > \rho^{\log \zeta_1 n/(1+\epsilon)\log n}$; or if $\mu\sqrt{\log_{1/p_1}|\mathcal{L}_n^0|} > \sqrt{2(1+\delta_2)\log n}$ for some δ_2, then we have

$$\mathbb{P}(\text{accept } \mathbb{H}_1(\mathcal{L}_n^0, \mu) \big| \mathbb{H}_0) + \mathbb{P}(\text{accept } \mathbb{H}_0 \big| \mathbb{H}_1(\mathcal{L}_n^0, \mu)) \to 0, \quad \text{as} \quad n \to \infty, \tag{5.6.70}$$

which means our statistics is asymptotically powerful by Definition 5.2.1.

5.7 Numerical study

In this section, we implement the numerical study on the detection problem of the suspected chain with good continuation in a noisy image. For simplicity, in all the examples throughout this section, it is assumed to have $C = 1$ and $m = 10$.

5.7.1 $|\mathcal{L}_n^0| \sim O(n)$

In this subsection, we assume that $|\mathcal{L}_n^0| \geq \zeta_1 \cdot n$ for some unknown constant $1 > \zeta_1 > 0$. From Table 3, we can choose x^* to be the $90th$ percentile of the standard normal distribution. Thus $x^* = 1.2816$ and $\rho = 0.2691 < \frac{1}{3}$. In Figure 9 all the significant nodes are black i.e., $\{(i,j) \in \mathcal{S} : X(i,j) > x^*\}$, while non-significant nodes are white. Let $\epsilon = 0.0001$ and thus as shown in (5.5.67), μ should satisfy

Figure 9: Black nodes are significant while white nodes are not significant.

.

$$p_1 = \mathbb{P}(N(\mu,1) > x^*) = \mathbb{P}(N(\mu,1) > 1.2816) > \rho^{\log \zeta_1 n / 1.0001 \log n} \tag{5.7.71}$$

Table 4 gives value of μ which satisfies (5.7.71) according to different values of ζ_1 and n. As we can see when ζ_1 or the number of columns n increase, we have smaller value of μ in the table which indicates stronger detectability in the noisy image. Intuitively, the increasing length of the inhomogeneous chain under $\mathbb{H}_1(\mathcal{L}_n^0, \mu)$ yields strong visibility.

Below is the simulation result for $m = 10$, $n = 200$, $C = 1$ and $\zeta_1 = \frac{1}{10}$. See Figure 10.

When $\mu \leq 3$, by human eyes it is hard to tell if there is an embedded chain different from the background. However, our method works for $\mu > 1.2216$. Figure 11 gives

Table 4: The minimum detectability of μ when $C = 1$ and $m = 10$ and $|\mathcal{L}_n^0| = \zeta_1 n$.

ζ_1	1/10	1/5	1/4	1/3	1/2	1
n=200	1.2216	1.0307	0.9745	0.9052	0.8126	0.6661
n=300	1.1740	1.0017	0.9504	0.8869	0.8017	0.6661
n=500	1.1247	0.9710	0.9249	0.8675	0.7901	0.6661
n=1,000	1.0716	0.9375	0.8969	0.8461	0.7772	0.6661
n=2,000	1.0296	0.9107	0.8743	0.8288	0.7668	0.6661
n=5,000	0.9860	0.8824	0.8506	0.8105	0.7556	0.6661
n=10,000	0.9594	0.8650	0.8359	0.7991	0.7487	0.6661
n=100,000	0.8960	0.8232	0.8004	0.7716	0.7319	0.6661
n=1000,000	0.8553	0.7959	0.7772	0.7535	0.7207	0.6661

a simulation for $m = 10$, $n = 300$, $C = 1$ and with $\frac{1}{5}$ portion of nodes on a chain with good continuation. We can see in Figure 11 the chain becomes apparent when $\mu \geq 2.5$ and our theory supports the detectability of such a chain when $\mu > 1.1740$.

5.7.2 $|\mathcal{L}_n^0| = \zeta_1 \log n$

In this part, we assume that $|\mathcal{L}_n^0| = \zeta_1 \log n$ for some unknown constant $0 < \zeta_1 < 1/\log(2C + 1) = 0.9102$. Let $\eta = 0.0001$ and $|\mathcal{L}_n^0| = \zeta_1 \log n$ for which ζ_1 is

$$1/5, 1/4, 1/3, 1/2, 2/3, 3/4, 4/5.$$

Table 5 gives the minimum values of μ corresponding to different ζ_1 such that

$$\mu\sqrt{|\mathcal{L}_n^0|} > \sqrt{(2+\eta)\log(mn^2(2C+1))}.$$

The larger value of μ in Table 5 than that in Table 4 is because $\mathcal{L}_0^n$ in this part is much shorter than the previous one. We give the length of $\mathcal{L}_n^0$ in this subsection corresponding to different value of ζ_1 and n in Table 6. As we can see in Table 6, even if $n = 10^8$ and $\zeta_1 = \frac{4}{5}$, the inhomogeneous chain only consists of 15 pixels which is impossible to find by eyes unless with very large elevated mean.

Table 5: The minimum detectability of μ when $C = 1$, $m = 10$ and $|\mathcal{L}_n^0| = \zeta_1 \log n$

ζ_1	1/5	1/4	1/3	1/2	2/3	3/4	4/5
n=1,000	7.0602	6.1143	5.4688	4.4653	3.8670	3.6459	3.5301
n=10,000	6.8837	6.1569	5.3321	4.3536	3.7703	3.5547	3.4418
n=100,000	6.7755	6.0602	5.2483	4.2852	3.7111	3.4988	3.3877
n=1000,000	6.7025	5.9949	5.1917	4.2390	3.6711	3.4611	3.3512
n=10,000,000	6.6498	5.9477	5.1509	4.2057	3.6422	3.4339	3.3249
n=100,000,000	6.6100	5.9122	5.1201	4.1805	3.6204	3.4134	3.3050

Table 6: Length of $\mathcal{L}_n^0$ when $|\mathcal{L}_n^0| = \zeta_1 \log n$

ζ	1/5	1/4	1/3	1/2	2/3	3/4	4/5
n=1,000	2	2	3	4	5	6	6
n=10,000	2	3	4	5	7	7	8
n=100,000	3	3	4	6	8	9	10
n=1,000,000	3	4	5	7	10	11	12
n=10,000,000	4	5	6	9	11	13	13
n=100,000,000	4	5	7	10	13	14	15

5.7.3 $\zeta \log n < |\mathcal{L}_n^0| < Cn^{1-\delta}$

In this part, we consider the minimum detectability when $|\mathcal{L}_n^0|$ lies in between the above two subsections, i.e.,

$$\zeta \log n < \left|\mathcal{L}_n^0\right| \leq C \cdot n^{1-\delta}$$

for some $\delta > 0$ such as $|\mathcal{L}_n^0| = C_1\sqrt{n}$ and $|\mathcal{L}_n^0| = C_2 \log n$, where $C_2 > \zeta$. In both cases, the value of the elevated mean μ that can be detectable is within our expected range.

(a) $|\mathcal{L}_n^0| = c\sqrt{n}$. Let $p_1 = \mathbb{P}(N(\mu, 1) > x^*)$ and $\delta_2 = 0.0001$. Let μ be such that

$$\mu\sqrt{\log_{1/p_1}(c\sqrt{n})} > \sqrt{(2+\delta_2)\log mn}. \tag{5.7.72}$$

The minimum value of μ that satisfies (5.7.72) is listed in Table 7.

(b) $|\mathcal{L}_n^0| = c \log n$, where $c \geq \zeta$. Again let $p_1 = \mathbb{P}(N(\mu, 1) > x^*)$ and $\delta_2 = 0.0001$.

Table 7: The minimum detectability of μ when $m = 10$, $C = 1$ and $|\mathcal{L}_n^0| = c\sqrt{n}$

c	1/3	1/2	1	2	5	10	50
n=1,000	1.7438	1.6806	1.5921	1.5205	1.4433	1.3943	1.3918
n=10,000	1.6742	1.6309	1.5667	1.5120	1.4503	1.4098	1.3308
n=100,000	1.6340	1.6011	1.5508	1.5065	1.4551	1.4206	1.3516
n=1,000,000	1.6077	1.5812	1.5398	1.5025	1.4582	1.4285	1.3673
n=10,000,000	1.5891	1.5669	1.5317	1.4996	1.4611	1.4345	1.3795
n=100,000,000	1.5752	1.5561	1.5256	1.4974	1.4632	1.4393	1.3894

Let μ be such that

$$\mu\sqrt{\log_{1/p_1}(c\log n)} > \sqrt{(2+\delta_2)\log mn}. \tag{5.7.73}$$

We list the minimum value of μ that satisfies (5.7.73) in Table 8. In Table 7 and 8, we find that gradually the minimum detectable mean μ increases as n becomes larger. This is due to the fact that the ratio $|\mathcal{L}_n^0|/n$ becomes more and more negligible as n tends to ∞. Table 9 gives the ratio of the length $|\mathcal{L}_n^0|$ of the embedded chain to the column number n. When $n = 10^8$ and $c = 100$, the inhomogeneous chain only occupies about 1.8×10^5 portion of the images which is fairly negligible.

Table 8: The minimum detectability of μ when $m = 10$, $C = 1$ and $|\mathcal{L}_n^0| = c\log n$

c	1	2	5	10	50	100
n=1,000	1.8228	1.7005	1.5822	1.5123	1.3886	1.3463
n=10,000	1.8566	1.7480	1.6393	1.5738	1.4556	1.4145
n=100,000	1.8914	1.7913	1.6991	1.6266	1.5122	1.4721
n=1,000,000	1.9242	1.8304	1.7330	1.6727	1.5613	1.5219
n=10,000,000	1.9549	1.8657	1.7720	1.7135	1.6046	1.5659
n=100,000,000	1.9833	1.8977	1.8071	1.7501	1.6433	1.6051

5.8 Extension

In this section, we will discuss about the longest significant run approach (as in Section 5.5) in the detection problem of the m-by-n array of nodes $\mathcal{S}$ as $m \to \infty$

Table 9: The ratio of the length of the embedded chain to n.

c	1	2	5	10	50	100
n=1,000	6.9078e-3	1.3816e-2	3.4539e-2	6.9078e-2	3.4539e-1	6.9078e-1
n=10,000	9.2103e-4	1.8421e-3	4.6052e-3	9.2103e-3	4.6052e-2	9.2103e-2
n=100,000	1.1513e-4	2.3026e-4	5.7565e-4	1.1513e-3	5.7565e-3	1.1513e-2
n=1,000,000	1.3815e-5	2.7631e-5	6.9078e-5	1.3816e-4	6.9078e-4	1.3816e-3
n=10,000,000	1.1619e-6	3.2236e-6	8.0590e-6	1.6118-5	8.0590e-5	1.6118e-4
n=100,000,000	1.8421e-7	3.6841e-7	9.2103e-7	1.8421e-6	9.2103e-6	1.8421e-5

and $n \to \infty$. A model with similar structure is studied profoundly in [6]. After thresholding the values at the nodes with threshold x^*, under the null hypothesis $\mathbb{H}_0$, each node $(i, j) \in \mathcal{S}$ is significant with probability

$$p = \mathbb{P}(N(0, 1) > x^*).$$

Let $p_1 = \mathbb{P}(N(\mu, 1) > x^*)$ be the probability of significance under $\mathbb{H}_1(\mathcal{L}_n^0, \mu)$. We use $L_0(m, n)$ to denote the longest chain consisting of significant nodes only under $\mathbb{H}_0$ and $|L_0(m, n)|$ is length. In [53], the authors show that there exists a continuous function $\phi(p)$ which only depends on C and p, but not on m and n such that as $m \to \infty$ and $n \to \infty$, we have

$$\frac{|L_0(m, n)|}{\log(mn)} \to \frac{1}{\phi(p)}, \quad \text{in probability,} \tag{5.8.74}$$

for $p < p_c$. Besides, $\phi(p)$ is a strictly decreasing function and it is positive when $p < p_c$ and constantly 0 as $p \geq p_c$. In [53], it is shown that $p_c \geq \frac{1}{2C+1}$. Thus, we may choose x^* such that $p < \frac{1}{2C+1}$ under the null hypothesis. As m and n become sufficiently large, the length of the longest significant chain is at most $(1 + \epsilon)\frac{\log mn}{\phi(p)}$ for some $\epsilon > 0$. Given the above, we have the following revised detection algorithm for the case that $(m, n) \to (\infty, \infty)$. The first step is the same as in Section 5.6.

Detection Algorithms for $(m, n) \to (\infty, \infty)$**:**

- For a constant $\zeta < \frac{1}{\log(2C+1)}$, let $\mathcal{F}_n^t = \{\mathcal{L} \in \mathcal{F}_n : |\mathcal{L}| \leq \zeta \log n\}$. Take

$$X_t^* = \max_{\mathcal{L} \in \mathcal{F}_n^t} \sum_{(i,j) \in \mathcal{L}} \frac{X(i,j)}{\sqrt{|\mathcal{L}|}}.$$

  For any small $\eta > 0$, if $X_t^* > \tau_t^* = \sqrt{2(1+\eta)\log(mn^2)}$, reject $\mathbb{H}_0$; otherwise go to the next step.

- Take x^* such that $p = \mathbb{P}(N(0,1) > x^*) < \frac{1}{2C+1}$ to be the threshold of nodes to be significant. Let $\mathcal{E}_n = \{\mathcal{L} \in \mathcal{F}_n : Z(\mathcal{L}) = 1\}$. Find the longest chain $L_0(m,n)$ in $\mathcal{E}_n$. For small $\epsilon > 0$ if the length $|L_0(m,n)| > (1+\epsilon/2)\frac{\log(mn)}{\phi(p)}$, then reject $\mathbb{H}_0$; otherwise, go to the next step.

- Compute X_s^* as

$$\max_{\mathcal{L} \in \mathcal{E}_n} \sum_{(i,j) \in \mathcal{L}} \frac{X(i,j)}{\sqrt{|\mathcal{L}|}}.$$

  For small $\delta_2 > 0$, if $X_s^* > \sqrt{2(1+\delta_2)\log(mn)}$, then reject $\mathbb{H}_0$; otherwise accept $\mathbb{H}_0$.

As shown above, the first step takes $O(mn \log n)$ flops, the second $O(mn)$ and third $O(mn \log(mn))$. Hence this algorithm takes $O(mn \log(mn))$ flops in total with $O(mn \log(mn))$ required space for storage. Moreover, by the proofs in Section 5.5 together with Theorem 5.4.2, it is straightforward to see that under $\mathbb{H}_1(\mathcal{L}_n^0, \mu)$ if either$|\mathcal{L}_n^0| < \zeta \log n$ and $\mu\sqrt{|\mathcal{L}_n^0|} > \sqrt{2(1+\eta)\log mn^2}$, or $|\mathcal{L}_n^0| \geq \zeta_1 n$ for some $\zeta_1 > 0$ and $p_1 > \exp\{-\phi(p)\frac{\log \zeta_1 n}{(1+\epsilon)\log(mn)}\}$, or $\mu\sqrt{\log_{1/p_1}|\mathcal{L}_n^0|} > \sqrt{2(1+\delta_2)\log(mn)}$ for some δ_2, then we have

$$\mathbb{P}(\text{accept}\mathbb{H}_1(\mathcal{L}_n^0, \mu) \big| \mathbb{H}_0) + \mathbb{P}(\text{accept}\mathbb{H}_0 \big| \mathbb{H}_1(\mathcal{L}_n^0, \mu)) \to 0, \quad \text{as} \quad n \to \infty.$$

5.9 Significant Nodes for Multi-sensor Problem

In this part, we will discuss how to decide whether a node is significant for a multi-sensor problem. Suppose we have an integer lattice of size $[1,n] \times [1,m]$ where each

node has a value of either 0 or 1, i.e.,

$$X_{i,j} = 1 \quad \text{or} \quad 0, \text{where} \quad i = 1, \ldots, n; j = 1, \ldots, m.$$

However, for each particular (i, j) the value of the node $X_{i,j}$ is unknown. We observe K sensors whose value at location (i, j), denoted by $S^k_{i,j}$ for $k = 1, 2, \ldots, K$, is also 0 or 1 depending on $X_{i,j}$ such that for each $i = 1, \ldots, n$ and $j = 1, \ldots, m$,

- $\mathbb{P}(S^k_{i,j} = 1 \big| X_{i,j} = 1) = p_1 \quad \text{and} \quad \mathbb{P}(S^k_{i,j} = 0 \big| X_{i,j} = 1) = 1 - p_1;$
- $\mathbb{P}(S^k_{i,j} = 1 \big| X_{i,j} = 0) = p_2 \quad \text{and} \quad \mathbb{P}(S^k_{i,j} = 0 \big| X_{i,j} = 0) = 1 - p_2.$

We do not know the value of p_1 and p_2 but let us suppose that $0 < p_2 < p_1 < 1$ and that $p_2 \leq \frac{1}{2}$. In other words, this assumption tells us that the value on the sensor is more likely to be 1 if the corresponding value of the array, $X_{i,j}$, is 1. In this case, we may declare a node $X_{i,j}$ to be significant if the sum of all the values on the sensors at the corresponding location is no less than some threshold T, and insignificant otherwise. To be more precise, we have a estimator at (i, j), denoted by $Y_{i,j}$ such that

$$Y_{i,j} = \begin{cases} 1, & \text{if} \sum_{k=1,\ldots,K} S^k_{i,j} \geq T; \\ 0, & \text{if} \sum_{k=1,\ldots,K} S^k_{i,j} < T. \end{cases}$$

The following theorem shows that if $p_2 K < T < p_1 K$, then $Y_{i,j}$ tends to be $X_{i,j}$ in probability as K becomes sufficiently large.

Theorem 5.9.1. *Suppose there exists a constant* $0 < \lambda < 1$ *such that* $T = \lambda p_1 K + (1 - \lambda) p_2 K$*, where* T *is the membership threshold defined above. If the number of the sensors* $K \to \infty$*, then we have the following*

$$\mathbb{P}(Y_{i,j} \neq X_{i,j}) \to 0.$$

The proof of Theorem 5.9.1 is one application of Chernoff-Okamoto inequalities for binomial distribution (See [21]). In order to prove Theorem 5.9.1, we need to use the following lemma.

Lemma 5.9.2. *For binomial probabilities with trial number n and success probability p, probability of failure $1-p$, we define $B(k,n,p)$ to be the probability of at most k successes and $E(k,n,p)$ of at least k successes, then we have the following:*

$$B(k,n,p): = \sum_{0\le j\le k}\binom{n}{j}p^j q^{n-j} \le \exp\{-\frac{(np-k)^2}{2npq}\}, \text{if} \quad k\le np\le \frac{n}{2}, \tag{5.9.75}$$

$$E(k,n,p): = \sum_{k\le j\le n}\binom{n}{j}p^j q^{n-j} \le (\frac{np}{k})^k(\frac{nq}{n-k})^{n-k}, \text{if} \quad k\ge np. \tag{5.9.76}$$

Proof of Theorem 5.9.1. It is obvious that we have

$$\mathbb{P}(Y_{i,j}\neq X_{i,j}) = \begin{cases} \mathbb{P}(Y_{i,j}=1), & \text{if} \quad X_{i,j}=0 \\ \mathbb{P}(Y_{i,j}=0), & \text{if} \quad X_{i,j}=1 \end{cases}$$

in which the former is the type-I error and the latter is type-II error in the hypothesis testing

$$\mathbb{H}_0: X_{i,j}=0 \quad \text{vs} \quad \mathbb{H}_1: X_{i,j}=1.$$

It follows that under $\mathbb{H}_0$,

$$\mathbb{P}(Y_{i,j}=1) = \mathbb{P}(\sum_{1\le k\le K} S^k_{i,j}\ge T\big|X_{i,j}=0) = E(T,K,p_2), \tag{5.9.77}$$

and under $\mathbb{H}_1$ we have

$$\mathbb{P}(Y_{i,j}=0) = \mathbb{P}(\sum_{1\le k\le K} S^k_{i,j}\le T-1\big|X_{i,j}=1) = B(T-1,K,p_1). \tag{5.9.78}$$

If we set $T=\lambda p_1 K+(1-\lambda)p_2 K$, where $0<\lambda<1$, then by (5.9.75) and (5.9.76) we have

$$B(T-1,K,p_1)\le B(T,K,p_1)\le \exp(-K\frac{(1-\lambda)^2(p_1-p_2)^2}{2p_1(1-p_1)})\to 0, \quad \text{as} \quad K\to\infty,$$

and

$$E(T,K,p_2)\le ((\frac{p_2}{\lambda p_1+(1-\lambda)p_2})^{\lambda p_1+(1-\lambda)p_2}(\frac{1-p_2}{1-\lambda p_1-(1-\lambda)p_2})^{1-\lambda p_1-(1-\lambda)p_2})^K. \tag{5.9.79}$$

Thus, to show $E(T, K, p_2) \to 0$ as $K \to \infty$, it suffices to show

$$f(\lambda) := (\frac{p_2}{\lambda p_1 + (1-\lambda)p_2})^{\lambda p_1 + (1-\lambda)p_2} (\frac{1-p_2}{1-\lambda p_1 - (1-\lambda)p_2})^{1-\lambda p_1 - (1-\lambda)p_2} < 1,$$

for any $\lambda \in (0, 1)$. It is obvious that $f(0) = 1$ and we will see that $\frac{d}{d\lambda} \log(f(\lambda)) < 1$. Indeed,

$$\begin{aligned} \log(f(\lambda)) &= (\lambda p_1 + (1-\lambda)p_2)(\log(p_2) - \log(\lambda p_1 + (1-\lambda)p_2)) \\ &+ (1 - \lambda p_1 - (1-\lambda)p_2)(\log(1-p_2) - \log(1 - \lambda p_1 - (1-\lambda)p_2)), \end{aligned}$$

and

$$\frac{d}{d\lambda} \log(f(\lambda)) = (p_1 - p_2) \log(\frac{p_2(1 - \lambda p_1 - (1-\lambda)p_2)}{(\lambda p_1 + (1-\lambda)p_2)(1-p_2)}).$$

If $p_2 < p_1$, then it is easy to see that $\frac{d}{d\lambda} \log(f(\lambda)) < 0$ when $\lambda \in (0, 1)$. In other words, we know the function $f(\lambda)$ is strictly decreasing when λ is between 0 and 1. □

Remark 5.9.3. *In this theorem, we see that either the probability of committing type-I error under $\mathbb{H}_0$ or type-II error under $\mathbb{H}_1$ has an exponential decay as K increases. Therefore, we do not need a very large K to guarantee the estimation error within accuracy of some small positive number.*

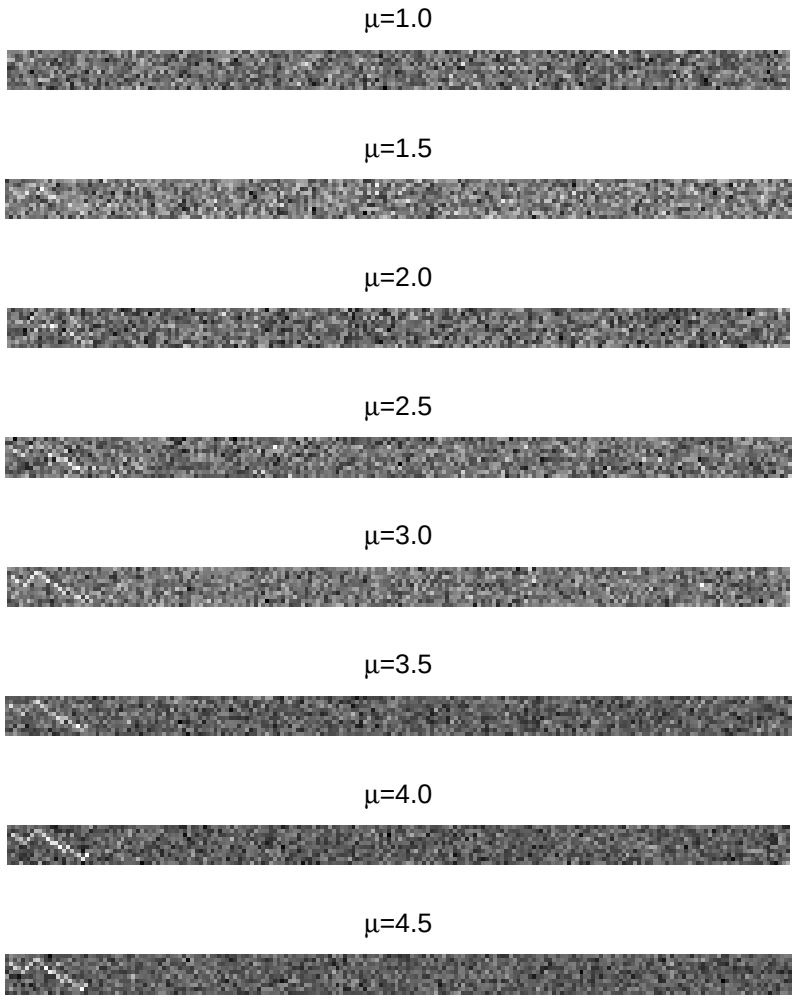

Figure 10: Grayscale images of 10×200 pixels with different means under $\mathbb{H}_1$ for a chain of length 20. When the elevated mean is less than 3.0, it is very hard to identify the inhomogeneous chain.

.

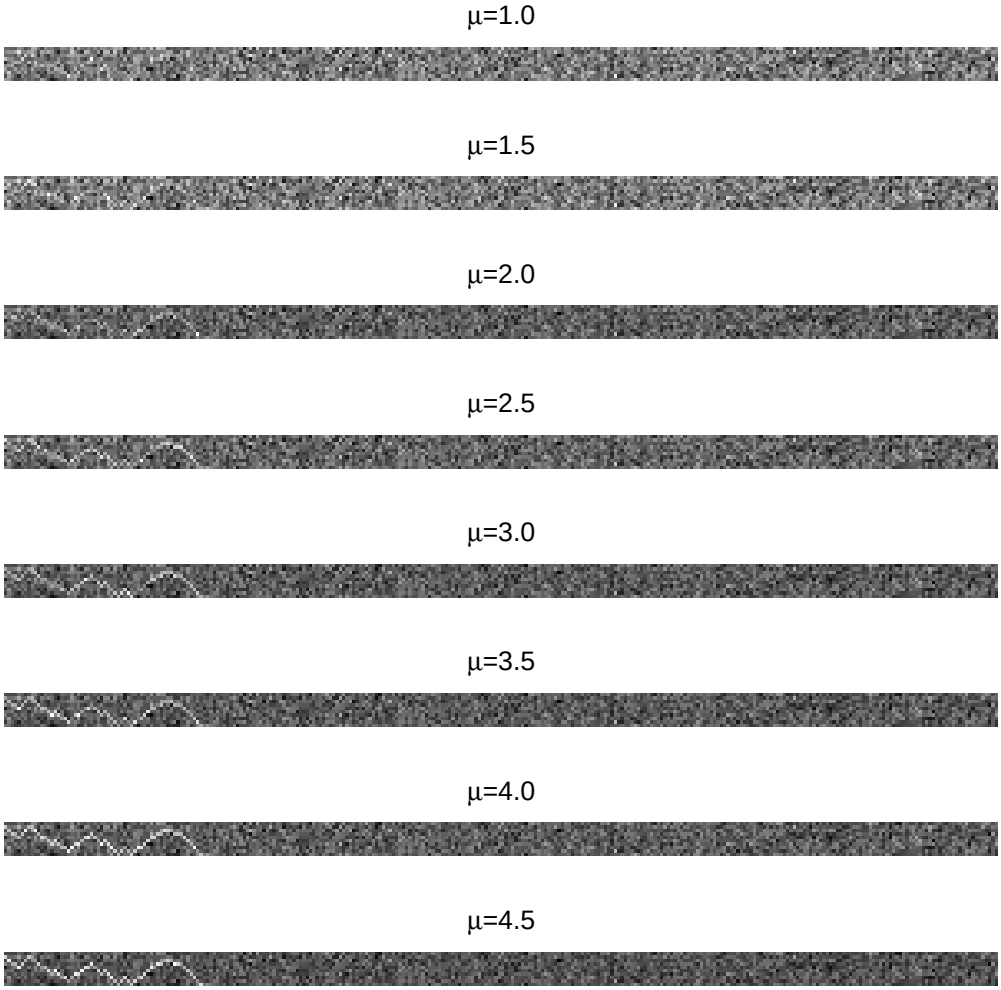

Figure 11: Gray-scale images of 10×300 pixels with different means under $\mathbb{H}_1$ for a chain of length 60. The inhomogeneous chain with good continuation becomes apparent when $\mu = 2.5$.

.

CHAPTER VI

CONCLUSION AND FUTURE WORK

In this thesis, we develop the asymptotic rate of the length of the longest significant run in an inflating Bernoulli net as $m \to \infty$ and $n \to \infty$. We also find the upper and lower bounds of the asymptotic probability to have a run with length k in the Pseudo-tree model. We apply our theory to the image detection problem to find the reasonable thresholds, which yields a reliable detection. It is of interests to learn the value of the function $\phi(p)$ in the future. Also for the portion of the nodes in the suspiciously curve, ϵ_N, we develop a lower bound, which guarantees a reliable test. However, it is our future work to find the minimum bound of $\epsilon_N > 0$, below which there is no powerful statistical test. Also it is not easy to find $\rho(m, p)$ when $m \geq 12$, especially when we drop the independence assumption among the nodes within the same column.

We also give a detection method for good continuous chains with elevated means in a white noise image. The numeric study shows the results are very promising, compared to human eyes' detectability. However, since the value of $\phi(p)$ is unknown yet and the value of $\rho(m, p)$ is hard to obtain when m is large, we cannot provide numeric studies in this special case. But overall, our new method is a fast and efficient algorithm as shown in simulations.

REFERENCES

[1] A. Hobolth, J. P. and Jensen, E. B. V., "A deformable template model, with special reference to elliptical templates," *J. Math. Imaging Version*, vol. 17, no. 2, pp. 131–137, 2002.

[2] A. Jain, Y. Z. and Dubuisson-Jolly, M., "Deformable template models:," *A review. Signal Processing*, vol. 71, no. 2, pp. 109–129, 1998.

[3] Arias-Castro, E., "Finite size percolation in regular trees," *Statistics and Probability Letters*, vol. 81, pp. 302–309, Feb. 2011.

[4] Arias-Castro, E., Donoho, D. L., and Huo, X., "Adaptive multiscale detection of filamentary structures embedded in a background of uniform random points," *Annals of Statistics*, vol. 34, pp. 326–349, Feb. 2006.

[5] Arias-Castro, E., Donoho, D., and Huo, X., "Near-optimal detection of geometric objects by fast multiscale methods," *IEEE Transcations on Information Theory*, vol. 51, pp. 2402–2425, July 2005.

[6] Arias-Castro, E. and Grimmett, G. R., "Cluster detection in networks using percolation," *Preprint*, Apr. 2011.

[7] Arratia, R., Goldstein, L., and Gordon, L., "Two moments suffice for poisson approximations: The chen-stein method," *The Annals of Probability*, vol. 17, pp. 9–25, Jan. 1989.

[8] Arratia, R., Gordon, L., and Waterman, M., "The erdös-r*énui* law in distribution, for coin tossing and sequence matching," *The Annals of Statistics*, vol. 18, pp. 539–570, 1990.

[9] Balakrishnan, N. and Koutras, M. V., "Runs and scans with applications," 2002.

[10] Bollobás, B. and Riordan, O., *Percolation*. Cambridge: Cambridge University PRess, 2006.

[11] Chen, J. and Huo, X., "Distribution of the length of the longest significance run on a bernoulli net, and its application," *Journal of American Statistical Association*, vol. 101, pp. 321–331, Mar. 2006.

[12] Cheng, S., Funke, S., Golin, M., Kumar, P., Poon, S., and Ramos, E., "Curve reconstruction from noisy samples," *Computational Geometry*, vol. 31, pp. 63–100, May 2005.

[13] Copeland, A. C., Ravichandran, G., and Trivedi, M., "Localized radon transform-based detection of ship wakes in sar image," *IEEE Trans. Geosci. Remote Sens.*, vol. 33, pp. 35–45, Jan. 1995.

[14] Council, N. R., *Expanding the vision of sensor materials*. Committee on New Sensor Technologies, Materials, and Applications, Washington, DC: National Academies Press, 1995.

[15] D. Culler, D. E. and Srivastava, M., "Overview of sensor networks," *IEEE Computers*, vol. 37, no. 8, pp. 41–49, 2004.

[16] D. Li, K. Wong, Y. H. H. and Sayeed, A., "Detection, classification, and tracking of targets," *Signal Processing Magazine, IEEE*, vol. 19, pp. 17–29, May 2002.

[17] D. Pozo, F. O. and Alados-Arboledas, L., "Fire detection and growth monitoring using a multitemporal technique on avhrr mid-infrared and thermal channels," *Remote Sensing of Environment*, vol. 60, no. 2, pp. 111–120, 1997.

[18] Davies, P., Langovoy, M., and Wittich, O., "Randomized algorithms for statistical image analysis based on percolation theory," tech. rep., submitted, 2009.

[19] Dey, T., *Curve and Surface Reconstruction: Algorithms with Mathematical Analysis*. Cambridge University Press, Mar. 2011.

[20] Duczmal, L. and Assuncao, R., "A simulated annealing strategy for the detection of arbitrary shaped spatial clusters," *Comput. Statist. Data Anal.*, vol. 45, no. 2, pp. 269–286, 2004.

[21] Dudley, R., *Uniform Central Limit Theorems*. Cambridge studies in advanced mathematics; 63, Cambridge University Press, 1999.

[22] Durrett, R., "Oriented percolation in two dimensions," *The Annals of Probability*, vol. 12, no. 4, pp. 999–1040, 1984.

[23] E. Arias-Castro, E. J. C. and Durand, A., "Detection of an anomalous cluster in a network," *The Annals of Statistics*, vol. 39, pp. 278–304, March 2011.

[24] E. Arias-Castro, E. J. Candes, H. H. and Zeitouni, O., "Searching for a trail of evidence in a maze," *The Annals of Statistics*, vol. 36, pp. 1726–1757, Aug. 2008.

[25] E. Arias-Castro, B. Efros, O. L., "Networks of polynomial pieces with application to the analysis of point clouds and images," *Journal of Approximation Theory*, pp. 94–130, Jan. 2010.

[26] Erdös, P. and Rényi, A., "On a new law of large numbers," *Journal of Analytical Mathematics*, vol. 22, pp. 103–111, 1970.

[27] Erdös, P. and Revesz, P., "On the length of the longest head run," *Colloqy of the Mathemtical Society of Janos Bolyai*, vol. 16, pp. 219–228, 1975.

[28] Esary, J., Proschan, F., and Walkup, D., "Association of random variables, with applications," *The Annals of Mathematical Statistics*, vol. 38, pp. 1466–1474, Oct. 1967.

[29] Fu, J., Wang, L., and Lou, W., "On exact and large deviation approximation for the distribution of the longest run in a sequence of two-state markov dependent trials," *Journal of Applied Probability*, vol. 40, pp. 346–360, 2003.

[30] G. P. Patil, J. Balbus, G. B. J. J. W. L. M. and Taillie, C., "Detection of an anomalous cluster in a network," *The Annals of Statistics*, vol. 39, pp. 278–304, March 2011.

[31] G. P. Patil, R. Modarres, W. L. M. and Patanker, P., "Spatially constrained clustering and upper level set scan hotspot detection in surveillance geoinformatics," *Environmental and Ecological Statistics*, vol. 13, no. 4, pp. 365–377, 2006.

[32] G. Patil, J. Balbus, G. B. J. J. W. M. and Taillie, C., "Multiscale advanced raster map analysis system: Definition, design and development," *Environmental and Ecological Statistics*, vol. 11, no. 2, pp. 113–138, 2004.

[33] G. Patil, S. J. and Koli, R., "Pulse, progressive upper level set scan statistic for geospatial hotspot detection," *Environmental and Ecological Statistics*, vol. 17, pp. 149–182, 2010.

[34] Geman, D. and Jedynak, B., "An active testing model for tracking roads in satellite images," *IEEE Trans Pattern Anal. March. Intell.*, vol. 18, pp. 1–14, 1996.

[35] Genovese, C., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L., "On the path density of a gradient field," *The annals of statistics*, vol. 37, pp. 3236–3271, 2009.

[36] Grimmett, G., *Percolation*. Grundlehren der mathematischen Wissenschaften, 321, Springer, 2nd ed., 1999.

[37] Hastie, T. and Stuetzle, W., "Principle curves," *Journal of American Statistical Association*, vol. 84, pp. 502–516, June 1989.

[38] Hills, R., "Searching for danger," *Science and Technology Review*, pp. 11–17, July 2001.

[39] Huo, X. and Chen, J., "Local linear projection," *First IEEE Workshop on Genomic Signal Processing and Statistics*, Dec. 2002.

[40] Huo, X. and Ni, S., "Detectability of convex-shaped objects in digital images, its fundamental limit and multiscale analysis," *Statistica Sinica*, vol. 19, pp. 1439–1462, Oct. 2009.

[41] Kegl, B., Krzyzak, A., Linder, T., and Zeger, K., "Learning and design of principal curves," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 281–297, Mar. 2000.

[42] Kulldorff, M., "A spatial scan statistic," *Comm. Statist. Theory Methods*, vol. 26, no. 6, pp. 1481–1496, 1997.

[43] Kulldorff, M., "Prospective time periodic geographical disease surveilance using a scan statistic," *J. Roy. Statist. Soc. Ser. A*, vol. 164, no. 1, pp. 61–72, 2001.

[44] Kulldorff, M. and Nagarwalla, N., "Spatial disease clusters: detection and inference," *Statistics in medicine*, vol. 14, no. 8, pp. 799–810, 1995.

[45] Langovoy, M., "Multiple testing, uncertainty and realisitic pictures," tech. rep., Technische Universiteit Eindhoven, EURANDOM, 2011.

[46] Langovoy, M. and Wittich, O., "Detection of objects in noisy images and site percolation on square lattices," tech. rep., Technische Universiteit Eindhoven, EURANDOM, 2009.

[47] Langovoy, M. and Wittich, O., "Robust nonparametric detection of objects in noisy images," tech. rep., Technische Universiteit Eindhoven, EURANDOM, 2010.

[48] Langovoy, M. and Wittich, O., "Multiple testing, uncertainty and realistic pictures," tech. rep., Technische Universiteit Eindhoven, EURANDOM, 2011.

[49] Lee, I.-K., "Curve reconstruction from unorganized points," *Computer Aided Geometric Design*, vol. 17, pp. 161–177, Sept. 1999.

[50] M. Kulldorff, L. Huang, L. P. and Duczmal, L., "An elliptic spatial scan statistic," *Stat. Med.*, vol. 25, no. 22, pp. 3929–3943, 2006.

[51] M. Kulldorff, Z. F. and Walsh, S. J., "A tree-based scan statistic for database disease surveillance," *Biometrics*, vol. 59, no. 2, pp. 323–331, 2003.

[52] McInerney, T. and Terzopoulos, D., "Deformable models in medical image analysis: a survey," *Medical Image Analysis*, vol. 1, no. 2, pp. 91–108, 1996.

[53] Ni, K. and Huo, X., "Asymptotic convergence rate of the lonest run in a bernoulli net," tech. rep., Georgia Institute of Technology, Atlanta, GA, 2012.

[54] Novikov, D., Colombi, S., and Dore, O., "Skeleton as a probe of the cosmic web: the 2d case," *Mon. Not. R. Astron. Soc.*, vol. 366, pp. 1201–1216, Feb. 2008.

[55] Papastavridis, S. G. and Koutras, M. V., "Bounds for reliability of consecutive-k-within-m-out-of-n systems,," *IEEE Transactions on Reliability*, vol. 42, pp. 156–160, 1993.

[56] Petrov, V., "On the probabilities of large deviations for sums of independent random variables," *Theory of Probability and Its Applications*, vol. 10, pp. 287–298, 1965.

[57] Pollard, D., *Convergence of Stochastic Processes*. Springer-Verlag, New York, 1984.

[58] R. Heffernan, F. Mostashari, D. D. A. K. M. K. and Weiss, D., "Syndromic surveillance in public health practice, new york city," *Emerging infectious diseases*, vol. 10, no. 8, pp. 858–864, 2004.

[59] Reid, D., "An algorithm for tracking multiple targets," *IEEE Transactions on Automatic Control*, vol. AC-24, pp. 843–854, Dec. 1979.

[60] Rotz, L. and Hughes, J., "Advances in detecting and responding to threats from bioterrorism and emerging infectious disease," *Nature Medicine*, pp. S130–S136, 2004.

[61] Roweis, S. and Saul, L., "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, Dec. 2000.

[62] Royden, H. and Fitzpatrick, P., *Real Analysis*. Pearson, 4th ed., 2010.

[63] Royden, H., *Real Analysis*. Prentice Hall, 3rd ed., 1988.

[64] S. M. Brennan, A. M. Mielke, D. C. T. and Maccabe, A. B., "Radiation detection with distributed sensor networks," *IEEE Computer*, vol. 37, pp. 57–59, 2004.

[65] Sandilya, S. and Kulkarni, S., "Principal curves with bounded turn," *IEEE Transactions on Information Theory*, vol. 48, pp. 2789–2793, Oct. 2002.

[66] Smola, A., Mika, S., Schoelkopf, B., and Williamson, R., "Regularized principle manifolds," *The journal of machine learning research*, vol. 56, pp. 459–477, Aug. 2007.

[67] Stoica, R., Martinez, V., and Saar, E., "A three-dimensional object point process for detection of cosmic filaments," *Journal of royal statistical society: Series C*, vol. 1, pp. 179–209, Sept. 2001.

[68] Szor, P., *The Art of Computer Virus Research and Defense*. Addison-Wesley Professional, 2005.

[69] Tango, T. and Takahashi, K., "A flexibly shaped spatial scan statistic for detecting clusters," *International Journal of Health Geographics*, vol. 4, no. 1, p. 11, 2005.

[70] Tenenbaum, J., de Silva, V., and Langford, J., "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, Dec. 2000.

[71] Tibshirani, R., "Principal curves revisited," *Journal of statistics and computing*, vol. 2, pp. 183–190, 1992.

[72] Y. Cui, Q. Wei, H. P. and Lieber, C., "Nanowire nanosensors for highly sensitive and selective detection of biological and chemical species," *Science*, vol. 293, pp. 1289–1292, 2001.

VITA

Kai Ni was born in Hangzhou, China on February 19, 1984. In September 2002, he went to Zhejiang University in Hangzhou, China and graduated with a Bachelor of Science in Mathematics in June, 2006. In August of 2007, he went to Atlanta and joined Georgia Institute of Technology to work in image detection and applied probabilistic problems under the supervision of Professors Vladimir Koltchinskii and Xiaoming Huo. Kai has earned master degrees in Statistics and Industrial Engineering at School of Industrial and System Engineering, Georgia Tech.

On November 28, 2010 in Minnesota, he married Cui Sun, who graduated from University of Georgia and is currently a Chinese immersion teacher at Minnetonka Public Schools, MN.